APPENDIX B

SELECTION AND FITTING OF DISTRIBUTIONS

B.0 INTRODUCTION

An important step in Monte Carlo analysis (MCA) is to select the most appropriate distributions to represent the factors that have a strong influence on the risk estimates. This step in the development of a Monte Carlo model can be very challenging and resource intensive.

☞ *Specifying probability distributions for all of the input variables and parameters in a probabilistic risk assessment (PRA) will generally not be necessary.*

If the sensitivity analysis indicates that a particular input variable does not contribute significantly to the overall variability and uncertainty, then this variable may be represented as a point estimate. As discussed in Appendix A, however, different approaches to sensitivity analysis may be applied throughout the tiered approach (e.g., sensitivity ratios, correlation analysis), and the ability to reliably identify variables as being minor or major can vary. Sometimes it can be helpful to develop probability distributions based on preliminary information that is available from Tier 1 in order to explore alternative options for characterizing variability and uncertainty. Likewise, sometimes the important "risk drivers" are apparent, and resources can be allocated to fully characterize the variability and uncertainty in those input variables. Therefore, the process of selecting and fitting distributions may also be viewed as a tiered approach. This appendix reviews the methods available to select and fit distributions and provides guidance on the process for determining appropriate choices depending on the information needed from the assessment and the information available to define the input variables.

In PRA, there are some important distinctions in the terminology used to describe probability distributions. A probability density function (PDF), sometimes referred to as a probability model, characterizes the probability of each value occurring from a range of possible values. Probability distributions may be used to characterize variability (PDFv) or uncertainty (PDFu). One advantage of using a PDFv and PDFu is that distributions represent a large set of data values in a compact way (Law and Kelton, 1991). For example, a lognormal distribution provides a good fit to a large data set of tap water ingestion rates (n=5,600) among children ages 1 to 11 years (Roseberry and Burmaster, 1992). Therefore, the distribution type (lognormal) and associated parameters (mean and standard deviation) fully describes the PDFv for intake rates, from which other statistics of interest can be calculated (e.g., median, and 95th percentile). Reducing a complex exposure model to a series of representative and well-fitting distributions can facilitate both the quantitative analysis and the communication of the modeling methodology. Alternatively, a PDFu may be specified to characterize parameter uncertainty. For example, the sample mean ($\bar{x}$) is generally an uncertain estimate of the population mean (μ) due to measurement error, small sample sizes, and other issues regarding representativeness (see Section B.3.1). A PDFu can be used to represent the distribution of possible values for the true, but unknown parameter. Understanding whether uncertainty or variability is being represented by a PDF is critical to determining how the distribution and parameters should be specified and used in a PRA.

EXHIBIT B-1

DEFINITIONS FOR APPENDIX B

Bayesian Analysis - Statistical analysis that describes the probability of an event as the degree of belief or confidence that a person has, given some state of knowledge, that the event will occur. Bayesian Monte Carlo combines a prior probability distribution and a likelihood function to yield a posterior distribution (see Appendix D for examples). Also called subjective view of probability, in contrast to the frequentist view of probability.

Bin - Regarding a histogram or frequency distribution, an interval within the range of a random variable for which a count (or percentage) of the observations is made. The number of bins for a histogram is determined on a case-by-case basis. In general, equal interval widths are used for each bin; however, in some cases (e.g., Chi-square test), individual bin widths are calculated so as to divide the distribution into intervals of equal probability.

Countably Infinite - Used to describe some discrete random variables, this term refers to a set of numbers that can be counted with integers (e.g., one, two, three) and that has no upper limit. Examples include the number of tosses required for a coin to show a head—we can count each toss, but it is possible that at least one more toss is needed. The number of dust particles in a volume of air is another example. Countably finite implies there is an upper limit (e.g., days of work per year).

Cumulative Distribution Function (CDF) - Obtained by integrating the PDF, gives the cumulative probability of occurrence for a random independent variable. Each value *c* of the function is the probability that a random observation *x* will be less than or equal to *c*.

Empirical Distribution Function (EDF) -The EDF, also called the empirical CDF (ECDF), is based on the frequency distribution of observed values for a random variable. It is a stepwise distribution function calculated directly from the sample, in which each data point is assigned an equal probability.

Frequency Distribution or Histogram - A graphic (plot) summarizing the frequency of the values observed or measured from a population. It conveys the range of values and the count (or proportion of the sample) that was observed across that range.

Goodness-of-Fit (GoF) Test - A method for examining how well (or poorly) a sample of data can be described by a hypothesized probability distribution for the population. Generally involves an hypothesis test in which the null hypothesis H_0 is that a random variable X follows a specific probability distribution F_0. That is, H_0: $F=F_0$ and H_a: $F \neq F_0$.

Independence - Two events *A* and *B* are independent if whether or not *A* occurs does not change the probability that *B* occurs. Likewise, knowing the value of *B* does not affect the value of *A*. Input variables, X and Y, are independent if the probability of any paired values (X, Y) is equal to the probability of X multiplied by the probability of Y. In mathematical terms, X and Y are independent if f(X, Y)=f(X) x f(Y). Independence is not synonymous with correlation. If X and Y are independent, then their correlation is zero, Cor(X, Y)= 0. But, the converse is not always true. There may be a nonlinear relationship between X and Y that yields Cor(X, Y)=0, but the variables are highly dependent.

Nonparametric Method - Also called a *distribution-free* method, a procedure for making statistical inferences without assuming that the population distribution fits a theoretical distribution such as normal or lognormal. Common examples are the Spearman rank correlation, (see Appendix A) and the bootstrap-t approach..

Parameter - In PRA, a parameter is a quantity that characterizes the probability distribution of a random variable. For example, a normal probability distribution may be defined by two parameters (e.g., arithmetic mean and standard deviation).

Parametric Distribution - A theoretical distribution specified by a distribution type and one or more parameters. Examples include the normal, Poisson, and beta distributions.

EXHIBIT B-1 —*Continued*
DEFINITIONS FOR APPENDIX B

Probability Density Function (PDF) - A function representing the probability distribution of a continuous random variable. The density at a point refers to the probability that the variable will have a value in a narrow range about that point.

Probability Distribution - The mathematical description of a function that associates probabilities with specified intervals or values for a random variable. A probability distribution can be displayed in a graph (e.g., PDF or CDF), summarized in a table that gives the distribution name and parameters, or expressed as a mathematical equation. In PRA, the process of selecting or fitting a distribution that characterizes variability or uncertainty can also be referred to as applying a *probability model* to characterize variability or uncertainty. In this guidance, the probability model is considered to be one source of model uncertainty.

Step Function - A mathematical function that remains constant within an interval, but may change in value from one interval to the next. Cumulative distribution functions for discrete random variables are step functions.

Z-score - The value of a normally distributed random variable that has been standardized to have a mean of zero and a SD of one by the transformation $Z=(X-\mu)/\sigma$. Statistical tables typically give the area to the left of the z-score value. For example, the area to the left of z=1.645 is 0.95. Z-scores indicate the direction (+/-) and number of standard deviations away from the mean that a particular datum lies assuming X is normally distributed. Microsoft Excel's *NORMSDIST(z)* function gives the probability p such that $p=\Pr(Z \leq z)$, while the *NORMSINV(p)* function gives the z-score z_p associated with probability p such that $p=\Pr(Z \leq z_p)$.

B.1.0 CONCEPTUAL APPROACH FOR INCORPORATING A PROBABILITY DISTRIBUTION IN A PRA

EXHIBIT B-2

GENERAL STRATEGY FOR SELECTING AND FITTING DISTRIBUTIONS

(1) Hypothesize a family of distributions
(2) Assess quality of fit of distribution
(3) Estimate distribution parameters
(4) Assess quality of fit of parameters

Often, more than one probability distribution may appear to be suitable for characterizing a random variable. A step-wise, tiered approach is recommended for incorporating probability distributions in a PRA. This appendix provides guidance on selecting and fitting distributions for variability and parameter uncertainty based on the overall strategy given in Exhibit B-2. Many of the same principles of selecting and fitting distributions are also given in EPA's *Report of the Workshop on Selecting Input Distributions for Probabilistic Assessments* (U.S. EPA, 1999a).

Probability distributions may be developed to characterize variability or uncertainty. Example flow charts for specifying a PDFv and PDFu are given in Figures B-1 and B-2, respectively. Both approaches outline an iterative process that involves three general activities: (1) identify potentially important sources of variability or uncertainty to determine if a PDF may be needed; (2) apply the general strategy given in Exhibit B-1 and evaluate plausible alternatives for distributions and parameter estimates; and (3) document the decision process. The flowcharts provide a general outline of the process and contain terms which are explained in subsequent sections. Just as with the point estimate approach, different sites may require different probability distributions for input variables, depending on the unique risk management issues and sources of uncertainty.

B.2.0 PRELIMINARY SENSITIVITY ANALYSIS

Selecting and fitting probability distributions for *all* of the input variables can be resource intensive and is generally unnecessary. Ideally, a subset of variables could be identified that contribute to most of the variability and uncertainty in a risk estimate. Sensitivity analysis can play an important role in helping to identify and quantitatively rank the major exposure pathways and variables. Since the information obtained from a sensitivity analysis may vary, depending on the approach(es) used and the information available to characterize the input variables, risk assessors should understand inherent limitations of each approach. A variety of approaches that are common for Tier 1 and 2 analyses are described and applied to a hypothetical example in Appendix A.

In a Tier 1 assessment, sensitivity analysis is typically limited to exploring the effect of alternative point estimates on the risk estimate. These methods can be helpful if additional information regarding the variability in the input variables is incorporated into the analysis (i.e., sensitivity scores). Alternatively, a reasonable approach is to specify preliminary probability distributions for one or more inputs in order to maximize the advantages of probabilistic methods. The difference between a preliminary distribution and a subsequent distribution reflects the level of effort invested in characterizing variability and uncertainty. If a robust data set is available in Tier 1 to define point estimates, then a preliminary distribution may, in fact, fully characterize variability with very high confidence. For other variables, summary statistics, rather than sample data, may be available, allowing for estimates of central tendency or plausible ranges. The use of preliminary distributions reflects an effort to employ more robust sensitivity analysis techniques without expending the effort and resources that might otherwise be applied to a PRA in Tier 2. The goal of the preliminary analysis would not be necessarily to evaluate risks and/or develop a PRG; rather, the focus would be on identifying input variables that may be important to explore more fully. Preliminary sensitivity analysis can provide insight into the importance of selecting among alternative probability distributions and exposure scenarios.

One-dimensional Monte Carlo simulations with preliminary (or screening-level) distributions can be run prior to engaging in a more involved process of selecting and fitting distributions. The distributions can be selected based on knowledge regarding the mechanisms that result in variability, and information already available for determining point estimates (e.g., summary statistics, U.S. EPA guidance, etc.). Table B-1 provides examples of preliminary distributions that might be selected based on the type of information available, sometimes referred to as the *state of knowledge*. In many cases, the distribution is intended to estimate the plausible bounds of a variable, while requiring no additional data collection effort. For example, given estimates of a lower bound [min], upper bound [max], and the assumption that each value is equally likely, a uniform distribution would be used to represent variability (or parameter uncertainty). If no mechanistic basis for selecting a distribution exists, then the preliminary distribution would be chosen based on the available information. For example, given the estimates of the arithmetic mean [μ] and a percentile value [a] for a random variable, an exponential distribution might be recommended with $\lambda=1/\mu$.

Guidance on matching the choice of the distribution to the state of knowledge is extended to a more diverse array of scenarios later in this appendix (see Table B-4).

Table B-1. Examples of Preliminary Distributions Based on Information Available[1], [2]

Information / Constraints	**Distribution Shape**
[a, b]	uniform
[a, m, b]	triangular
[a, b, α_1, α_2, β]	beta
[μ, σ]	normal
γ	exponential
[a, b, μ, σ]	Johnson Sb, Lognormal
[α, β]	gamma

a=minimum, b=maximum, m=mode, α=shape parameter, μ=mean, σ=standard deviation, γ=average rate of occurrence of events, β=scale,

It may be informative to explore alternative choices for distributions applied to the same variable. For example, a simple yet informative approach is to run two 1-D MCA simulations for variability with an input variable characterized first by a Johnson Sb (i.e., a four-parameter lognormal distribution; Hahn and Shapiro, 1967) and then by a normal distribution. The difference in the risk distribution, especially at the percentile that is relevant to the risk management decision (e.g., 95th percentile), may offer insights regarding the importance of the shape of the PDFv.

B.3.0 WHAT DOES THE DISTRIBUTION REPRESENT?

Distributions may be specified to characterize variability or uncertainty. Often, a Monte Carlo simulation of variability will focus on describing differences between individuals in a population (i.e., inter-individual variability). In this case, the goal is to select a distribution that is representative of the *target* population—the set of all receptors that are potentially at risk. There may be uncertainty that the choice of PDFv reflects variability in the target population. In general, risk assessors should fully disclose uncertainties in the PDFv, especially because the use of a distribution instead of a point estimate may inappropriately suggest that there is a greater state of knowledge. Following the tiered process (see Chapter 2, Figure 2-1), there are multiple opportunities to consider consequences of alternative modeling approaches early in the process of developing a probabilistic model. The importance of relating the distribution to the *target* population, clearly distinguishing between variability and uncertainty, and evaluating data representativeness is emphasized in Sections B.3.1, B.3.2 and B.4.

[1] The preliminary distributions are based in part on maximum entropy concepts. Maximum entropy is a technique for determining the distribution that represents the maximum uncertainty allowed by the available information and data (Vose, 1996). Although the approach can be used to quickly define distributions that maximize uncertainty, the credibility of the distribution depends on the use of accurate, unbiased information.

[2] See Table B-2 for more detailed descriptions of selected distributions.

B.3.1 CONCEPTS OF POPULATION AND SAMPLING

The distinction between a *target* population, a *sampled* population, and a *statistical* population should be considered carefully when evaluating information for use in both Tier 1 and Tier 2 of a PRA. The *target* population is often considered to be the “population of concern”. A risk assessor is often interested in quantifying specific attributes of the population (e.g., exposure duration, exposure frequency, etc.). A *sampled* population is the set of receptors available for selection and measurement. For purposes of this appendix/guidance, the *sampled* population may be the *target* population or it may be a different population that is thought to be representative of the *target* population. For purposes of this guidance, a *statistical* population is an approximation of the *target* population based on information obtained from the *sampled* population.

Distributions are generated from representative *sample* populations to make inferences about the *target* population. Ideally, a *sampled* population should be a subset of a *target* population and should be selected for measurement to provide accurate and representative information about the exposure factor being studied. However, defining representative samples is a matter of interpretation.

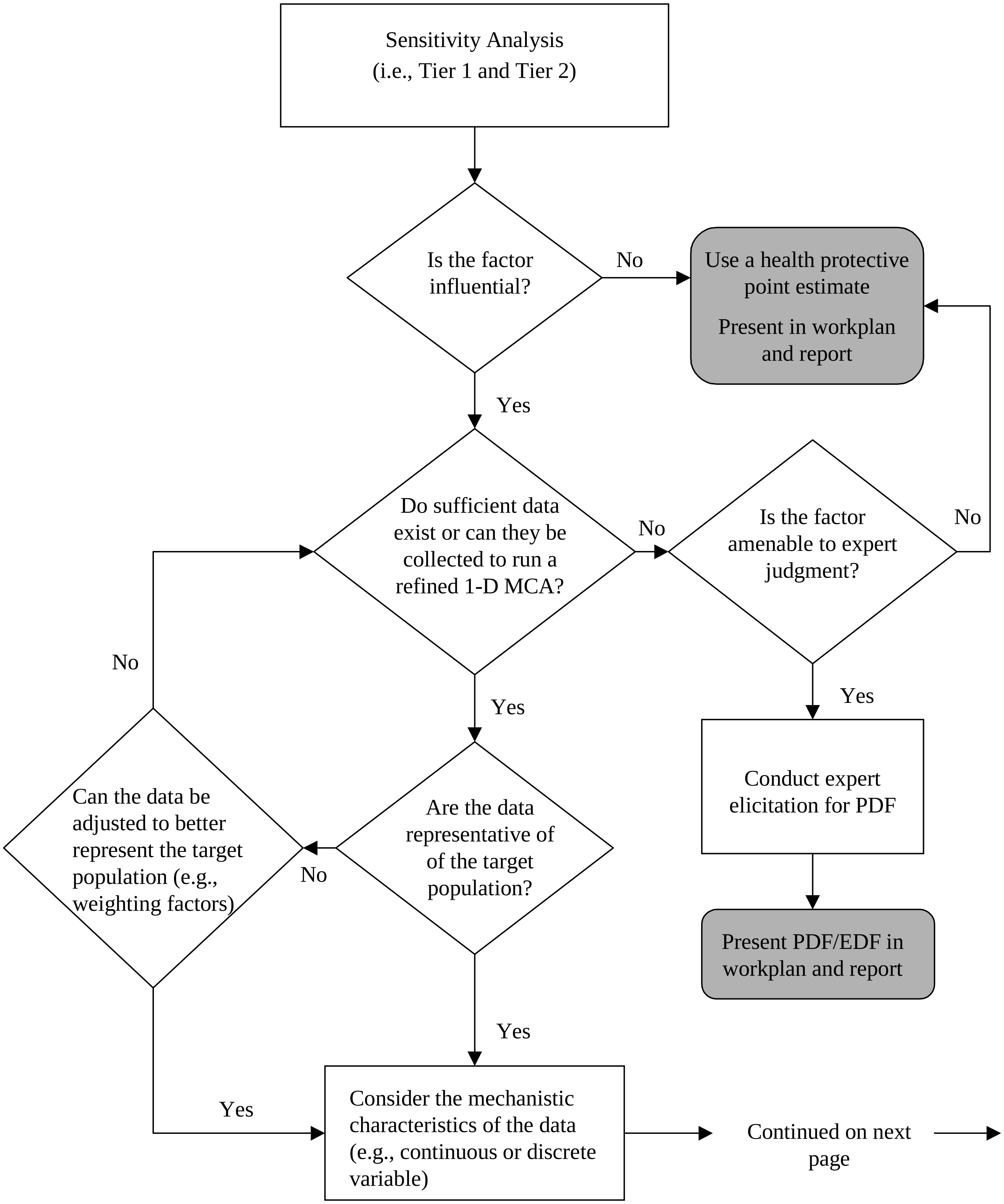

Sensitivity Analysis
(i.e., Tier 1 and Tier 2)
Is the factor influential?
No
Use a health protective point estimate
Present in workplan and report
Yes
Do sufficient data exist or can they be collected to run a refined 1-D MCA?
No
Is the factor amenable to expert judgment?
No
No
Yes
Yes
Conduct expert elicitation for PDF
Can the data be adjusted to better represent the target population (e.g., weighting factors)
Are the data representative of of the target population?
No
Present PDF/EDF in workplan and report
Yes
Yes
Consider the mechanistic characteristics of the data (e.g., continuous or discrete variable)
Continued on next page

Figure B-1 (page 1 of 2). Conceptual approach for incorporating probability distributions for variability in PRA.

(Continued from previous page)

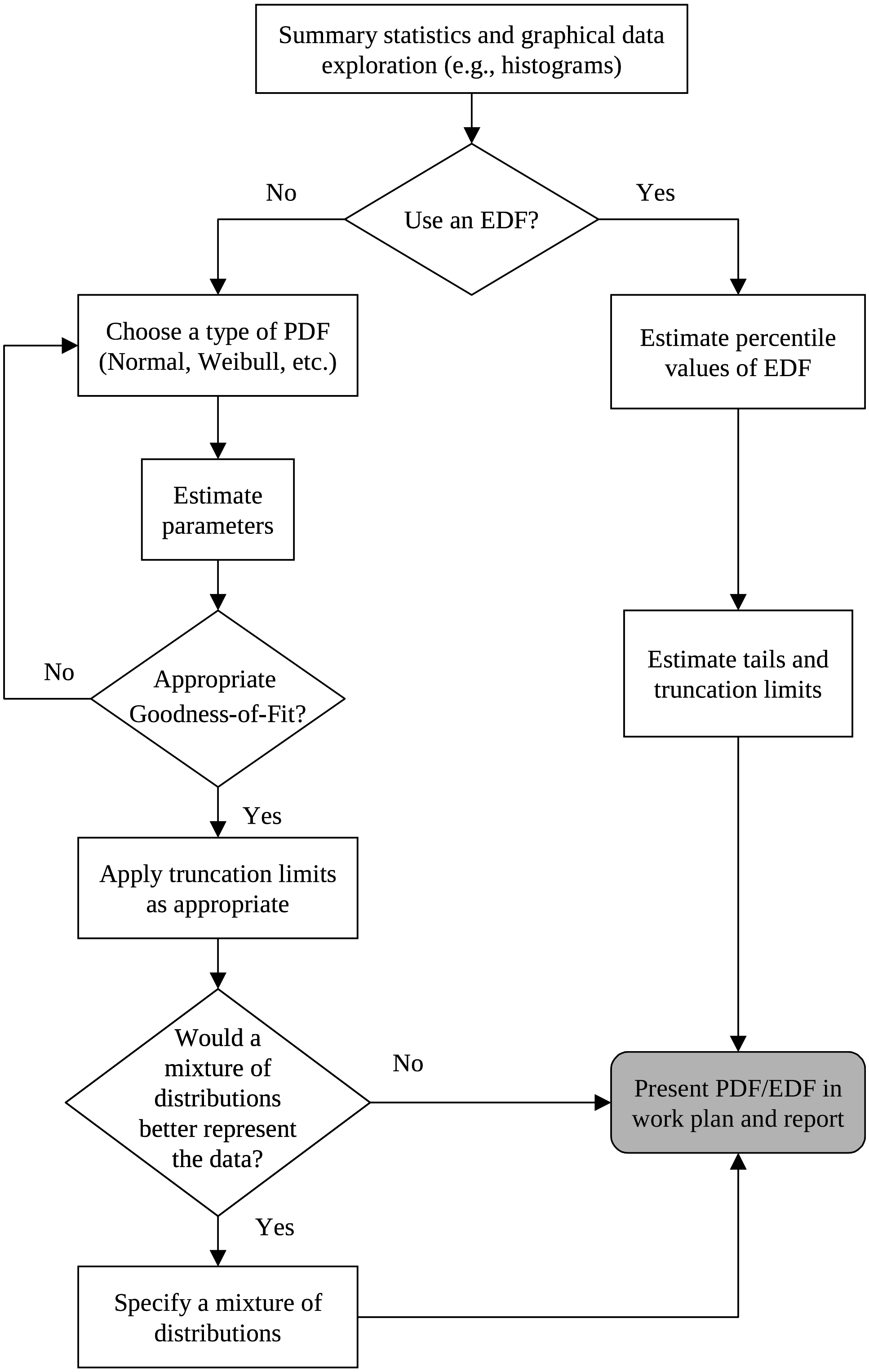

Summary statistics and graphical data exploration (e.g., histograms)
Use an EDF?
No
Yes
Choose a type of PDF (Normal, Weibull, etc.)
Estimate percentile values of EDF
Estimate parameters
Appropriate Goodness-of-Fit?
No
Estimate tails and truncation limits
Yes
Apply truncation limits as appropriate
Would a mixture of distributions better represent the data?
No
Present PDF/EDF in work plan and report
Yes
Specify a mixture of distributions

Figure B-1 (page 2 of 2). Conceptual approach for incorporating probability distributions for variability in PRA.

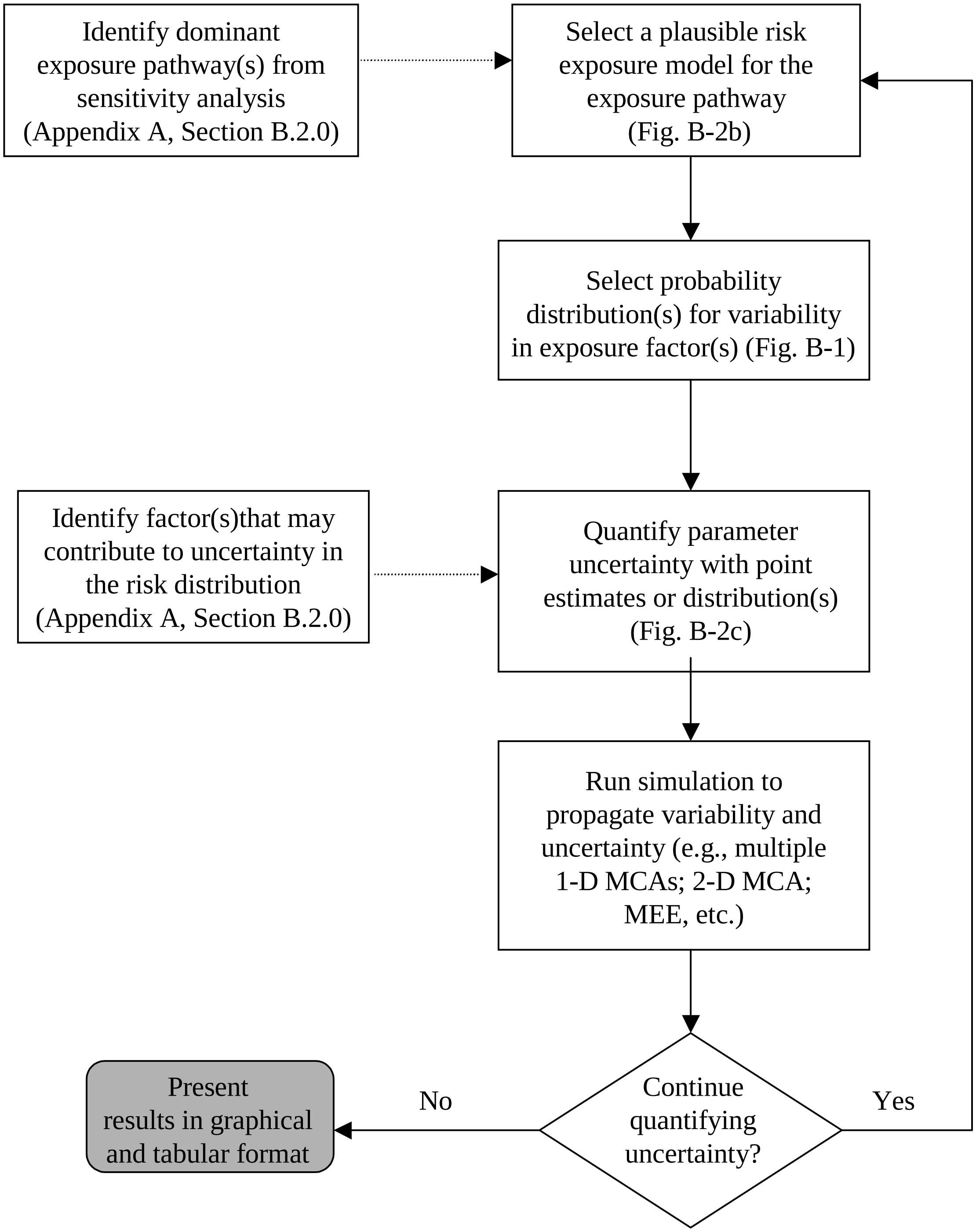

Identify dominant exposure pathway(s) from sensitivity analysis (Appendix A, Section B.2.0)
Select a plausible risk exposure model for the exposure pathway (Fig. B-2b)
Select probability distribution(s) for variability in exposure factor(s) (Fig. B-1)
Identify factor(s)that may contribute to uncertainty in the risk distribution (Appendix A, Section B.2.0)
Quantify parameter uncertainty with point estimates or distribution(s) (Fig. B-2c)
Run simulation to propagate variability and uncertainty (e.g., multiple 1-D MCAs; 2-D MCA; MEE, etc.)
Present results in graphical and tabular format
No
Continue quantifying uncertainty?
Yes

Figure B-2a (page 1 of 3). Conceptual approach for quantifying model and parameter uncertainty in PRA.

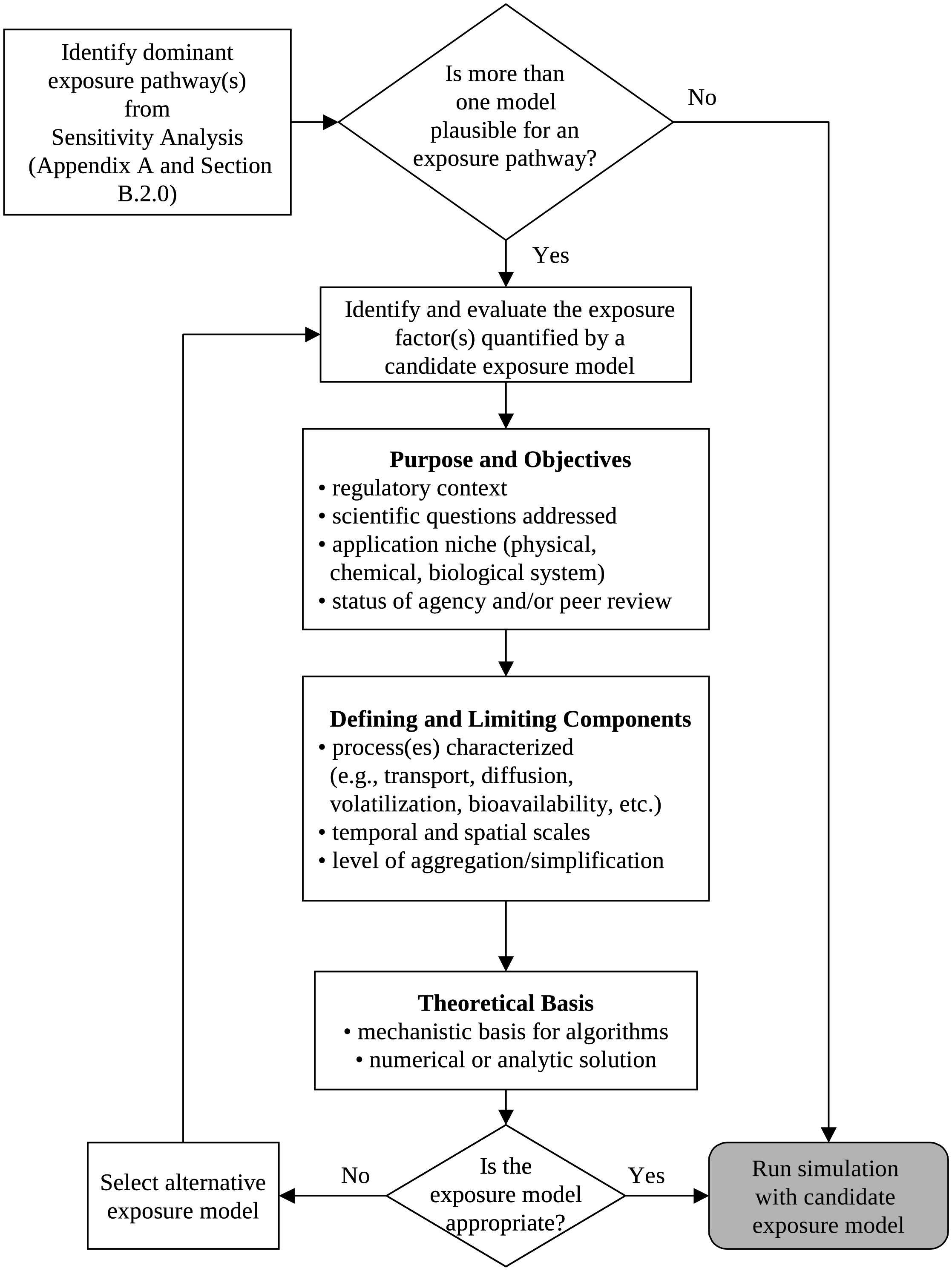

Identify dominant exposure pathway(s) from Sensitivity Analysis (Appendix A and Section B.2.0)
Is more than one model plausible for an exposure pathway?
No
Yes
Identify and evaluate the exposure factor(s) quantified by a candidate exposure model
Purpose and Objectives
• regulatory context
• scientific questions addressed
• application niche (physical, chemical, biological system)
• status of agency and/or peer review
Defining and Limiting Components
• process(es) characterized (e.g., transport, diffusion, volatilization, bioavailability, etc.)
• temporal and spatial scales
• level of aggregation/simplification
Theoretical Basis
• mechanistic basis for algorithms
• numerical or analytic solution
Select alternative exposure model
No
Is the exposure model appropriate?
Yes
Run simulation with candidate exposure model

Figure B-2b (page 2 of 3). Detailed conceptual approach for incorporating model uncertainty in PRA.

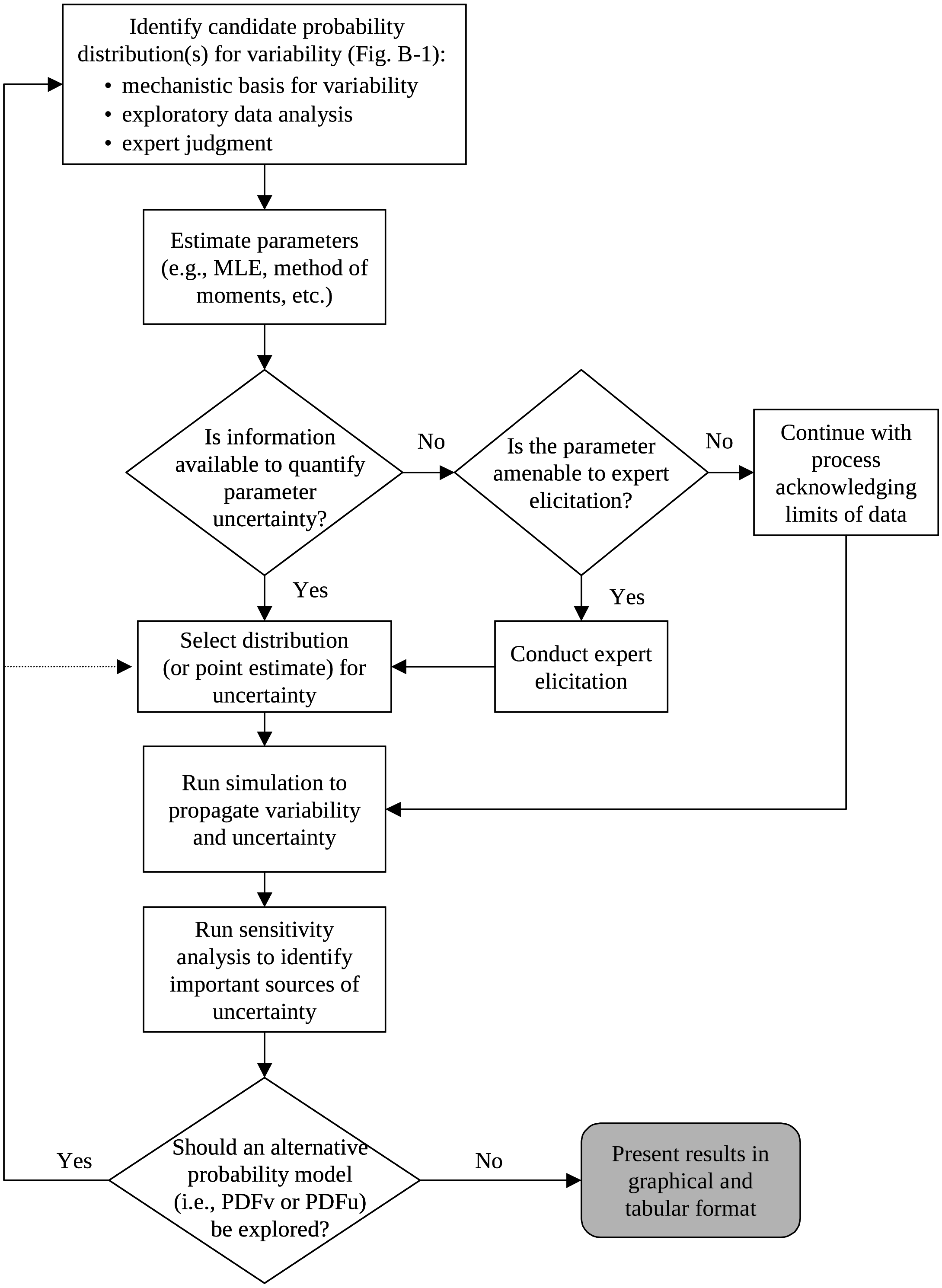

Identify candidate probability distribution(s) for variability (Fig. B-1):
• mechanistic basis for variability
• exploratory data analysis
• expert judgment
Estimate parameters (e.g., MLE, method of moments, etc.)
Is information available to quantify parameter uncertainty?
No
Is the parameter amenable to expert elicitation?
No
Continue with process acknowledging limits of data
Yes
Yes
Select distribution (or point estimate) for uncertainty
Conduct expert elicitation
Run simulation to propagate variability and uncertainty
Run sensitivity analysis to identify important sources of uncertainty
Yes
Should an alternative probability model (i.e., PDFv or PDFu) be explored?
No
Present results in graphical and tabular format

Figure B-2c (page 3 of 3). Detailed conceptual approach for incorporating parameter uncertainty in PRA.

B.3.2 CONSIDERING VARIABILITY AND UNCERTAINTY IN SELECTING AND FITTING DISTRIBUTIONS

Multiple probability distributions may be used to describe variability and uncertainty in an input variable. For example, a normal probability distribution may be selected to characterize variability in body weight, whereas a uniform distribution may selected to characterize uncertainty in the estimate of the arithmetic mean of the normal distribution. The appropriate interpretation and analysis of data for an exposure variable will depend on whether one is specifying a PDFv or PDFu. Figure B-1 outlines one useful process for selecting distributions for variability, whereas Figure B-2 (three pages) outlines a useful process for quantifying both model and parameter uncertainty.

Variability generally refers to observed differences attributable to true heterogeneity or diversity in a population (U.S. EPA, 1997b). Variability results from natural random processes. Inter-individual variability may stem from environmental, lifestyle, and genetic differences. Examples include human physiological variation (e.g., natural variation in body weight, height, breathing rates, drinking water intake rates), changes in weather, variation in soil types, and differences in contaminant concentrations in the environment. Intra-individual variability may reflect age-specific changes (e.g., body weight and height). Variability is not reducible by further measurement or study. A PDF for variability can usually be obtained by fitting a distribution to the sample measurements.

Sources of Uncertainty

Uncertainty generally refers to the lack of knowledge about specific factors, parameters, or models (U.S. EPA, 1997b). Although uncertainty in exposure and risk assessment may be unavoidable due to the necessary simplification of real-world processes, it generally can be reduced by further measurement and study. Parameter uncertainty may stem in part from measurement errors, sampling errors, or other systematic errors in the collection and aggregation of data. Model uncertainty may reflect the simplification of a complex process, a mis-specification of the exposure model structure, a misuse or misapplication of an exposure model, use of the wrong distributional model, and the use of surrogate data or variables. Scenario uncertainty may reflect uncertainty in an exposure model, such as the relevance of specific exposure pathways to the target population. A conceptual exposure model can be used to provide direction in specifying a probability distribution for uncertainty. For example, the concentration term in a Superfund risk assessment typically represents the long-term average concentration to which a receptor is exposed (see Chapter 5). An uncertainty distribution for the concentration term could be developed in part from ideas about the statistical uncertainty of estimating the long-term average from a small sample, and the assumption of random movement of the receptors within a defined exposure unit.

Probability Distributions and Model Uncertainty

This appendix primarily focuses on methods for quantifying uncertainty associated with both the selection of a variability distribution, and estimating parameters of a distribution. A probability distribution can be referred to as a type of model in the sense that it is an approximation, and often a simplified representation of variability or uncertainty that combines both data and judgment. A broader use of the term model refers to a representation of a chemical, physical, or biological process. In risk assessment, many different models have been developed, with varying objectives, major defining and limiting components, and theoretical basis. Figure B-2b provides a general process for exploring model uncertainty of this type. This figure reflects the concepts and spirit of the *Agency Guidance for Conducting External Peer Review of Environmental Regulatory Modeling* (U.S. EPA, 1994). In general, EPA regional risk assessors should be consulted in order to determine the types of exposure and risk models that may be plausible for quantifying exposure at a particular site.

Parameter Uncertainty

Quantifying parameter uncertainty in a probabilistic model typically requires judgment (see Appendix C). When data are uncertain due to, for example, small sample sizes or questionable representativeness (Section B.3.1), Monte Carlo simulation can be a useful tool for demonstrating the effect of the uncertainty on the risk estimates. It is most important to model uncertainty when the sensitive input variables are uncertain. Uncertainty can be quantified in both the point estimate approach (e.g., a range of possible central tendency exposure values) or a probabilistic approach (e.g., a range of possible values for the arithmetic mean of a distribution). While a quantitative uncertainty analysis may complicate a risk management decision by suggesting that risk estimates are highly uncertain, this information can be helpful by focusing additional efforts towards collecting data and reducing uncertainty in the most sensitive input variables. Likewise, if an estimated risk is below a regulatory level of concern, even after quantifying highly uncertain inputs to the exposure model, the risk manager may be more confident in a decision. As emphasized in Figures B-2a, B-2b, and B-2c, risk assessors should generally refrain from setting *ad hoc* probabilities to different candidate distributions in a single Monte Carlo simulation. Instead, this guidance strongly recommends exploring model or parameter uncertainty by running a separate simulation with each candidate model. For example, rather than randomly assigning a beta distribution or a lognormal distribution to an exposure variable for each iteration of a simulation, separate simulations should be run with the candidate probability distributions. Similarly, if a range of temporal or spatial scales is plausible for quantifying exposure, multiple simulations should be designed to demonstrate the importance of these assumptions on the risk estimates.

Uncertainty in parameter estimates may be characterized using a variety of methods. Similar to a PDF for variability, a PDF for parameter uncertainty may be represented by a probability distribution with a unique set of parameters. Sometimes the distribution for uncertainty can be specified by knowing (or assuming) a distribution for variability. For example, if X is a normally distributed random variable, the Student's t distribution and the Chi-square (χ^2) distribution can be used to develop PDFu's for random measurement error uncertainty in the sample mean and variance, respectively. The PDFu for both the Student's *t* and Chi-square distributions is determined by the sample size (*n*). If a PDFu cannot be determined from the PDF for variability, or assumptions regarding the underlying distribution for variability are not supportable, nonparametric or "distribution free" techniques may be used (e.g., bootstrapping). Both parametric and nonparametric techniques may yield confidence intervals for estimates of population parameters.

B.4.0 DO DATA EXIST TO SELECT DISTRIBUTIONS?

Developing site-specific PDFs for every exposure assumption (or toxicity value, in the case of ecological risk) can be time and resource intensive, and in many cases, may not add value to the risk management decision. For those exposure variables that do exert a significant influence on risk, a PDF may be developed from site-specific data, data sets available in the open literature (e.g., EPA's *Exposure Factors Handbook*, U.S. EPA 1997a), or from existing PDFs in the literature (e.g., Oregon DEQ, 1998).

At Superfund sites, perhaps the most common exposure variable that will be described by site-specific data will be the media concentration term. The sample (i.e., collection of empirical measurements) will most often be used to estimate either a point estimate of uncertainty (e.g., an upper confidence limit for the arithmetic mean concentration—the 95% UCL), or a distribution that characterizes the full distribution of uncertainty in the mean. Exposure variables such as ingestion rates, exposure duration, and exposure frequency will most likely be derived from existing PDFs or data sets in the open literature. The Agency supports the development PDFs that may be generally applicable to different sites (e.g., body weight, water intake, and exposure duration) (U.S. EPA, 1999b, 2001). Until final recommendations of PDFs are available for the more generic exposure variables, PDFs for exposure variables that lack adequate site-specific data will typically be selected from: (1) existing PDFs; (2) data on the entire U.S. population; or (3) data on subsets of the U.S. population that most closely represent the target population at a site. If risks to a sensitive subpopulation, such as young children, elderly adults, ethnic groups, or subsistence fishermen, are a concern at a site, then existing PDFs or data sets that best characterize these subpopulations would be preferable to national distributions based on the entire U.S. population. If adequate site-specific data are available to characterize any of the exposure variables, distributions can be fit to those data.

Uncertainty Associated with Sample Size

An appropriate question to consider when evaluating data sets for use in exposure and risk assessment is, "What sample size is sufficient?" Generally, the larger the sample size (*n*), the greater one's confidence in the choice of a probability distribution and the corresponding parameter estimates. Conversely, for small *n*, Goodness-of-fit (GoF) tests (see Section B.6.2) will often fail to reject many of the hypothesized PDFs. In general, there is no rule of thumb for the minimum sample size needed to specify a distribution for variability or uncertainty. Increasing a sample size may be an appropriate option to consider when evaluating risk management strategies to reduce uncertainty.

Statistical sampling, in general, is important to consider when estimating parameters of a probability distribution. One rule of thumb is that the parameters that reflect the central tendency of a distribution (e.g., arithmetic mean, median, mode) can be estimated with greater confidence than parameters that reflect the extremes of the distribution (e.g., 95th percentile). When deciding on appropriate truncation limits (minimum and maximum values), it is unlikely that the statistical sample actually includes the plausible bounds. See Section B.5.7 for more detailed guidance on specifying truncation limits for probability distributions.

B.4.1 WHAT ARE REPRESENTATIVE DATA?

The question, "What is a representative sample?", is important to address when selecting and fitting distributions to data. Many of the factors that may determine representativeness (e.g., sample size and the method of selecting the target, and sample population (Section B.3.1)) are relevant to both point estimate and PRA. EPA's *Guidance for Data Usability in Risk Assessment, Part A* (U.S. EPA, 1992) describes representativeness for risk assessment as the extent to which data define the true risk to human health and the environment.

The goal of representativeness is easy to understand. However, evaluating data to determine if they are representative is more difficult, especially if the problem and decision objectives have not been clearly defined.

The importance of representativeness also varies with the level of complexity of the assessment. If a screening level assessment is desired, for example, to determine if concentrations exceed a health protective exposure level, then representativeness may not be as important as health protectiveness. However, if a complete baseline risk assessment is planned, the risk assessor should generally consider the value added by more complex analyses (e.g., site-specific data collection, sensitivity analysis, and exposure modeling). A tiered approach for making these decisions for a PRA is presented in Chapter 2, and examples of more complex analyses are presented in Appendix D. In addition, the Agency (U.S. EPA, 1999a) summarizes the advantages and weaknesses of proposed checklists for risk assessors to evaluate representativeness of exposure factors data.

For purposes of this guidance, a surrogate study is one conducted on a sampled population that is similar to, but not a subset of, the target population. When using surrogate data, the risk assessor should generally exercise judgment about the representativeness of the data to the target population. For example, the distribution of body weights of deer mice from two independent samples from similar ecosystems may differ depending on the age structure, proportion of males and females, and the time of year that the samples were obtained. When in doubt about which study results to use in defining a probability distribution, one option is to develop a distribution and calculate risks with each sample independently, and compare the results. This approach can be a simple, but effective type of uncertainty analysis. At a minimum, uncertainties associated with the use of surrogate studies should be discussed in the assessment.

In many cases, the surrogate population shares common attributes with the target population, but is not truly representative. The risk assessor should then determine the importance of the discrepancies and whether adjustments can be made to reduce those differences. There are a wide variety of methods that can be used to account for such discrepancies, depending on the available information. Summary statistics (e.g., as presented by the *Exposure Factors Handbook*, U.S. EPA, 1997a) can be used to estimate linear characteristics of the target population from the sample population. For example, if the mean, standard deviation, and various percentiles of the sample population are known, then the mean or proportion exceeding a fixed threshold can be calculated using a simple weighted average. Adjustment options are more numerous if the risk assessor has access to the raw data. Adjustments for raw data include: weighted averages, weighted proportions, transformations, and grouping of the data based on the available information (e.g., empirical data, and professional judgment).

In most cases, the evaluation of data representativeness will necessarily involve judgment. The workplan should generally include a description of the data, the basis for the selection of each distribution, and the method used to estimate parameters (see Chapter 2). Empirical data (i.e., observations) are typically used to select distributions and derive parameter estimates. However, it may be necessary to use expert judgment or elicitation in cases where the quality or quantity of available data are found to be inadequate.

B.4.2 THE ROLE OF EXPERT JUDGMENT

Expert judgment refers to inferential opinion of a specialist or group of specialists within an area of their expertise. When there is uncertainty associated with an input variable, such as a data gap, expert judgment may be appropriate for obtaining distributions. Note that distributions elicited from experts reflect individual or group inferences, rather than empirical evidence. Distributions based on expert judgment can serve as Bayesian priors in a decision-analytic framework. The distributions and Bayesian priors can be modified as new empirical data become available. There is a rich literature base regarding the protocol for conducting expert elicitations and using the results to support decisions (Morgan and Henrion, 1990). Elicitation of expert judgment has been used to obtain distributions for risk assessments (Morgan and Henrion, 1990; Hora, 1992; U.S. EPA, 1997b) and for developing air quality standards (U.S. EPA, 1982).

Bayesian analysis is a statistical approach that allows the current state of knowledge, expressed as a probability distribution, to be formally combined with new data to reach an updated information state. In PRA, Bayesian Monte Carlo analysis (Bayesian MCA) can be used to determine the reduction in uncertainty arising from new information. When combined with techniques from decision analysis, Bayesian MCA can help to determine the type and quantity of data that generally should be collected to reduce uncertainty. The benefits and limitations of expert elicitation, Bayesian statistics, Bayesian MCA, and decision analysis (i.e., value of information [VOI]), as applied to PRA, are discussed in greater detail in Appendix D.

EXHIBIT B-3

FACTORS TO CONSIDER IN SELECTING A PROBABILITY DISTRIBUTION*

- *Is there a mechanistic basis for choosing a distributional family?*
- *Is the shape of the distribution likely to be dictated by physical or biological properties or other mechanisms?*
- *Is the variable discrete or continuous?*
- *What are the bounds of the variable?*
- *Is the distribution skewed or symmetric?*
- *If the distribution is thought to be skewed, in which direction?*
- *What other aspects of the shape of the distribution are known*?
- *How well do the tails of the distribution represent the observations?*

*Source: U.S. EPA, 1997b

B.5.0 FITTING DISTRIBUTIONS TO DATA

Sometimes more than one probability distribution may adequately characterize variability or uncertainty. The choice of a distribution should be based on the available data and on knowledge of the mechanisms or processes that result in variability. In general, the preferred choice is the simplest probability model that adequately characterizes variability or uncertainty and is consistent with the mechanism underlying the data. For example, a log-logistic distribution would not necessarily be selected over a 2-parameter lognormal distribution simply because it was ranked higher in a GoF test by a statistical software package. Some distributions (e.g., normal, lognormal) are well known among risk assessors. The statistical properties for these distributions are well understood and the formal descriptions can often be brief.

Important factors to consider in selecting a PDF are described in Exhibit B-3. An initial step in selecting a distribution should be to determine if the random variable is discrete or continuous. Continuous variables take any value over one or more intervals and generally represent measurements (e.g., height, weight, concentration). For a continuous variable, a mathematical function generally describes the probability for each value across an interval. Discrete variables take either a finite or *countably infinite* number of values. Unique probabilities are assigned to each value of a discrete variable. The number of rainfall events in a month is an example of a discrete random variable, whereas the amount of rainfall is a continuous variable. Similarly, the number of fish meals per month is a discrete variable, whereas the average size (mass) of a fish meal is continuous.

Another important consideration is whether there are plausible bounds or limits for a variable. For example, it is highly unlikely that an American adult will weigh less than 30 kg or more than 180 kg. Most exposure variables may assume any nonnegative value within a plausible range. Therefore, distributions will generally be truncated at a minimum of zero (or higher), or a probability distribution that is theoretically bounded at a nonzero value may be specified (see Table B-3). A more detailed discussion of factors to consider in selecting a PDF and specifying parameter values is provided below.

B.5.1 CONSIDERING THE UNDERLYING MECHANISM

There may be mechanistic reasons depending on known physical or biological processes that dictate the shape of the distribution. For example, normal distributions result from processes that sum random variables whereas lognormal distributions result from multiplication of random variables. A Poisson distribution is used to characterize the number of independent and randomly distributed events in a unit of time or space. An exponential distribution would describe the inter-arrival times of independent and randomly distributed events occurring at a constant rate. If, instead, the elapsed time until arrival of the k^{th} event is of interest, then the appropriate probability distribution would be the gamma distribution (Morgan and Henrion, 1990).

☞ *In all cases, it is incumbent on the risk assessor to explain clearly and fully the reasoning underlying the choice of a distribution for a given exposure variable—primarily from a mechanistic standpoint if possible.*

Table B-2 lists some of the probability distributions that may commonly be used in PRA. This is not an exhaustive list, and the scientific literature contains numerous examples with alternative distributions. Where practicable, a mechanistic basis is presented for the choice of the distribution. For some distributions, such as beta, triangular, and uniform, a mechanistic basis is not offered because it is unlikely that a chemical or biological process will yield a random variable with that particular shape. Nevertheless, such distributions may be appropriate for use in PRA because they reflect the extent of information that is available to characterize a specific random variable. Preliminary distributions are discussed in Section B.2.0 and Table B-4. Because many of the distributions given in Table B-2 can assume flexible shapes, they offer practical choices for characterizing variability.

Table B-2 also illustrates probability distributions (both PDFs and CDFs) commonly used in PRA. While intuitively appealing, identifying a mechanistic basis for a distribution can be difficult for many exposure variables; however, it may be relatively apparent that the variable is bounded by a minimum (e.g., ingestion rate ≥ 0 mg/day) and a maximum (e.g., absorption fraction ≤ 100%), or that the relevant chance mechanism results in a discrete distribution rather than a continuous distribution, as described above.

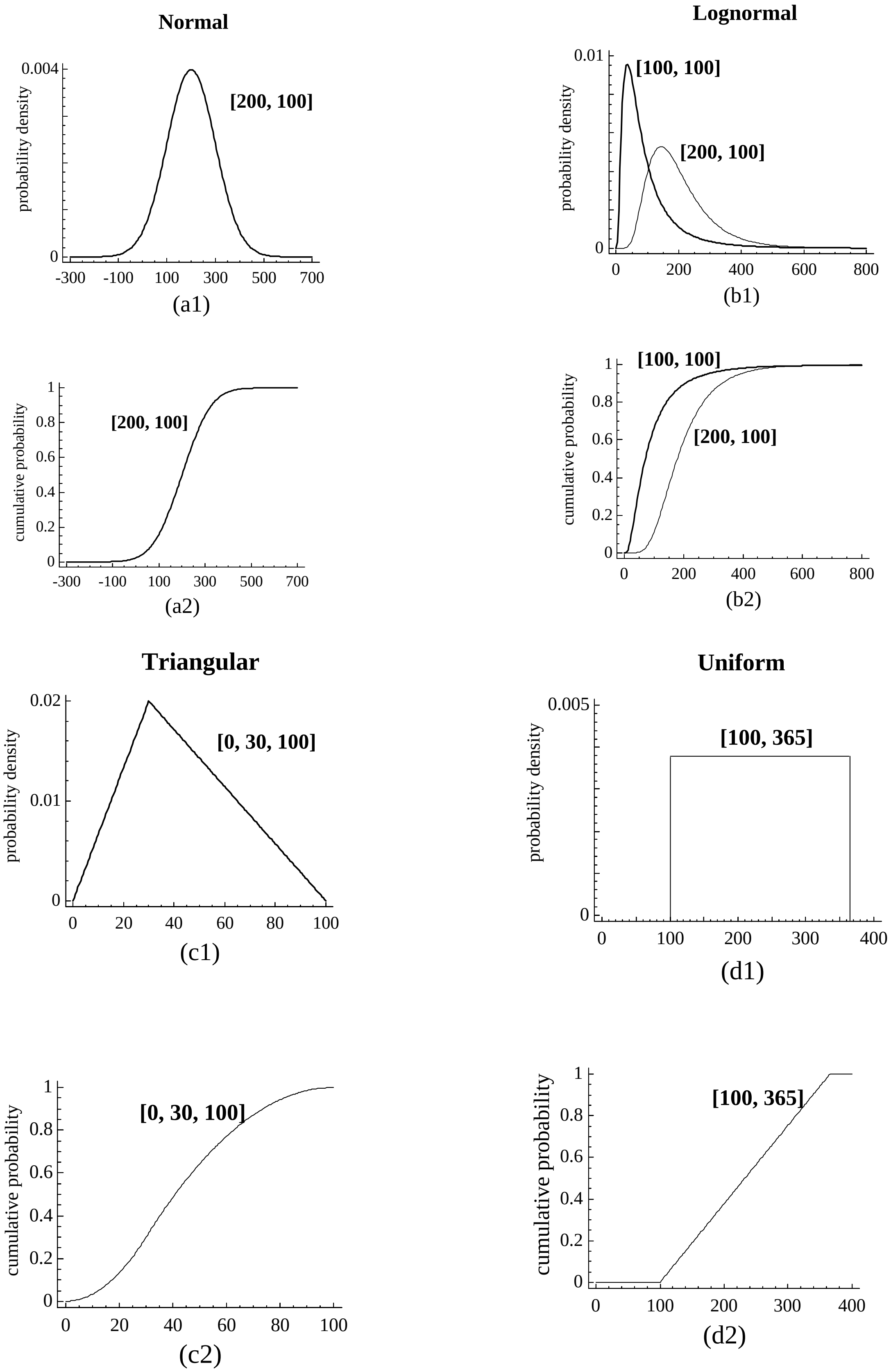
For each distribution, one or more examples with different parameter estimates are given to demonstrate the flexibility in the shape of the PDF. In addition to the descriptions of the distributions in Tables B-2, Table B-3 provides a summary of the parameters and theoretical bounds that define the PDFs. For a further discussion of characteristics of PDFs see Thompson, 1999. Figures (a-h) immediately following Table B-2 present examples of PDFs and the corresponding CDFs for distributions commonly used in PRA.

Table B-2. Examples of Selected Probability Distributions for PRA.

Distribution	Mechanistic Basis	Example(s)
Beta Figure (e)	Describes a continuous random variable with finite upper and lower bounds. This distribution can take on very flexible shapes, but generally does not have a mechanistic basis.	Absorption fraction bounded by 0 and 100%; fraction of time an individual spends indoors.
Binomial	Describes a discrete random variable produced by processes that: (1) occur in a fixed number *n* of repeated independent "trials"; (2) yield only one of two possible outcomes (e.g., "success" or "failure") at each trial; and (3) have constant probability *p* of "success". A binomial distribution is characterized by parameters *n*, *p*, and *x*, representing the number of trials, the probability of success of each trial, and the number of successes, respectively.	The number of animals with tumors (or some other quantitative outcome) in a chronic animal bioassay.
Exponential Figure (h)	If instead of counting the number of events in the Poisson process (below), one measures the time (or distance) between any two successive, random, independent events.	The length of time between two radiation counts; length of time between major storm events; distance between impact points of two artillery shells.
Gamma Figure (g)	Similar to exponential except that time until occurrence of the k^{th} event in the Poisson process is measured (rather than time between successive events). Reduces to exponential when k=1.	Time until k^{th} radiation count; elapsed time until k^{th} major storm event.
Lognormal Figure (b)	Multiplication of a large number of random variables, or equivalently adding the logarithms of those numbers, will tend to yield a distribution with a lognormal shape.	Chemical concentrations in environmental media; media contact rates; rates and flows in both fate and transport models. Because the basic risk equation is multiplicative, distributions of risk are generally lognormal. In practice, lognormal distributions often provide good fits to data on chemical concentrations in a variety of media (Gilbert, 1987; Ott, 1990).
Normal Figure (a)	Addition of independent random variables, with no one variable contributing substantially to the total variation of the sum, will tend to yield a distribution with a normal shape. This result is established by the central limit theorem.	The "Gaussian Plume Model" for the dispersion of air pollutants is based on the idea that, at a micro level, individual parcels of air, or molecules of pollutants, are subject to many random collisions from other molecules that act together as if a large number of random numbers were being added/subtracted from an initial 3-dimensional description of a position.

Table B-2. Examples of Selected Probability Distributions for PRA.

Distribution	Mechanistic Basis	Example(s)
Poisson	Observed when counting the frequency of discrete events, where the events are independent of one another, and randomly distributed in space or time. Approximates the binomial distribution when sample size, *n*, is large and probability, *p*, is small.	The number of counts of radiation that occur in a particular time interval; the release of synaptic transmitter from nerve cells; the number of artillery shells falling within a fixed radius; the occurrence of major storm events in a month; number of leaks in average length of pipe.
Triangular Figure ©)	The PDF is shaped like a triangle, with parameters representing plausible bounds and a most likely value (i.e., mode). This is a “rough” probability model that generally describes a random variable based on limited information rather than mechanistic basis.	Variability in shower droplet diameter. Uncertainty in the mean air exchange rate in a shower.
Uniform Figure (d)	The PDF is shaped like a rectangle, with parameters representing plausible bounds. This is a “rough” probability model that generally describes a random variable based on limited information rather than a mechanistic basis.	Variability in the air ventilation rate in a house.
Weibull Figure (f)	Originated in reliability and (product) life testing as a model for time to failure or life length of a component when the failure rate changes with time. A very flexible model taking a wide range of shapes. If the failure rate is constant with time, the Weibull reduces to the exponential distribution.	Examples for exponential and gamma would also be appropriate for Weibull.



(a1)

(b1)

(a2)

(b2)

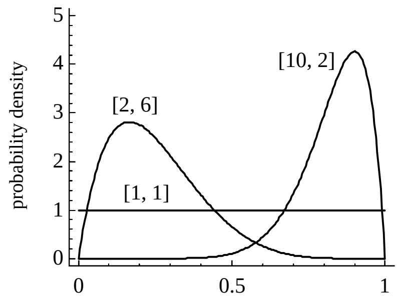
(c1)

(d1)

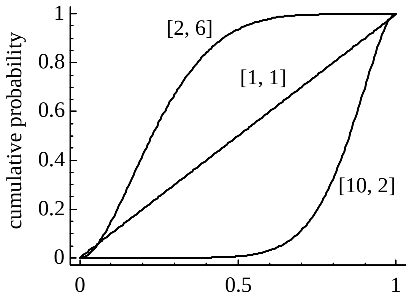
(c2)

(d2)

Beta

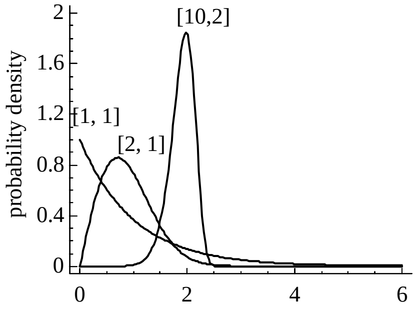


(e1)

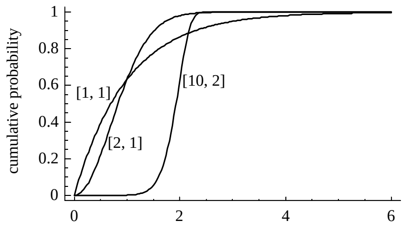


(e2)

Weibull

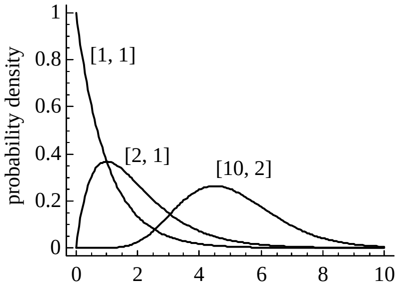


(f1)

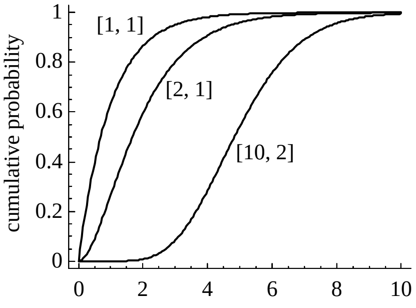


(f2)

Gamma

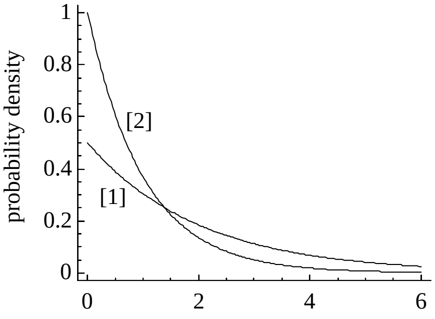


(g1)

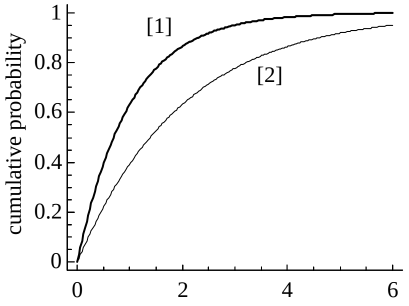


(g2)

Exponential



(h1)



(h2)

B. 5.2 EMPIRICAL DISTRIBUTION FUNCTIONS (EDFS)

In some cases, an empirical distribution function (EDF) may be preferred over fitting the data set to a hypothesized distribution. EDFs, also called empirical cumulative distribution functions (ECDF), provide a way to use the data itself to define the distribution of the relevant variable. Briefly, an EDF for a random variable is described by a step function based on the frequency distribution of observed values. An EDF for a continuous random variable may be linearized by interpolating between levels of the various bins in a frequency distribution. The CDF for a linearized EDF appears as a line, rather than steps. Example B-3 at the end of this Appendix illustrates an EDF, linearized EDF, and beta distribution (α_1=0.63, α_2=2.85, rescaled to min=0, max=364) fit to percentile data for soil ingestion rates in children (Stanek and Calabrese, 1995). A plausible range (i.e., minimum and maximum values) was imposed on the data set for this example.

EDFs provide a complete representation of the data with no loss of information. They do not depend on the assumptions associated with estimating parameters for theoretical probability models. EDFs are designed to provide direct information about the shape of the distribution, which reveals skewness, multimodality, and other features of the data set. However, EDFs may not adequately represent the tails of a distribution due to limitations in data acquisition. In the simplest case, an EDF is constrained to the extremes of the data set. This may be an unreasonable restriction if limiting the EDF to the smallest and largest sample values is likely to greatly underestimate the distributional tails. If this is an important source of uncertainty, the risk assessor may choose to extend the tails of the distribution to plausible bounds or to describe the tails with another distribution (see Exhibit B-4). For example, an exponential distribution may be used to extend the tails based on the last 5% of the data. This method is based on extreme value theory, and the observation that extreme values for many continuous, unbounded distributions follow an exponential distribution (Bratley et al., 1987). As with other probability models, uncertainty in the plausible bounds of an EDF may be reduced by obtaining additional information.

EXHIBIT B-4

VARIATIONS OF THE EDF

Linearized - Linearly interpolates between two observations, yielding a linearized cumulative distribution pattern.

Extended - In addition to linearizing (see above), adds lower and upper bounds based on expert judgment.

Mixed Exponential - Adds an exponential upper and/or lower tail to the EDF.

Advantages and disadvantages of using EDFs in PRA are discussed in detail in the *Report of the Workshop on Selecting Input Distributions for Probabilistic Assessments* (U.S. EPA, 1999a).

B.5.3 GRAPHICAL METHODS FOR SELECTING PROBABILITY DISTRIBUTIONS

Graphical methods can provide valuable insights and generally should be used in conjunction with exploratory data analysis. Examples of graphical methods are frequency distributions (i.e., histograms), stem-and-leaf plots, dot plots, line plots for discrete distributions, box-and-whisker plots, and scatter plots (Tukey, 1977; Conover, 1980; Morgan and Henrion, 1990).

☞ *Graphical methods are invaluable for exploring a data set to understand the characteristics of the underlying population.*

Together with statistical summaries, graphical data summaries can reveal important characteristics of a data set, including skewness (asymmetry), number of peaks (multi-modality), behavior in the tails, and data outliers.

Frequency Distribution or Histogram

The frequency distribution, or histogram, is a graphical approximation of the empirical PDF. Frequency distributions can be plotted on both linear and log scales. The general strategy for selecting the number of bins to partition the data is to avoid too much smoothing and too much jaggedness. Equation B-1 (U.S. EPA, 1999a) provides a starting point for estimating the number of bins based on the sample size (*n*).

$$Number\,of\ Bins = 1 + 3.322\log_{10} n \qquad \text{Equation B-1}$$

Probability Plotting

Another method that may be used to visualize distributions and estimate parameters is probability plotting, also referred to as linear least squares regression or regression on ordered statistics. This technique involves finding a probability and data scale that plots the CDF of a hypothesized distribution as a straight line. The corresponding linearity of the CDF for the sample data provides a measure of the GoF of the hypothesized distribution. The general approach involves sorting the sample data in ascending order and converting the ranks to percentiles. The percentile value for the i^{th} rank is calculated according to Gilbert (1987) as:

$$Percentile = 100 \times \frac{i - 0.5}{n} \qquad \text{Equation B-2}$$

An alternative formula is provided by Ott (1995):

$$Percentile = 100 \times \frac{i}{n + 1} \qquad \text{Equation B-3}$$

Plotting positions given by Equations B-2 and B-3 are special cases of the more general formula given by Equation B-4 (Helsel and Hirsch, 1992):

$$Percentile = 100 \times \frac{i - a}{n + 1 - 2a} \qquad \text{Equation B-4}$$

where *a* is a constant that varies from 0 (Equation B-3) to 0.5 (Equation B-2).

The percentiles are used to calculate the *z-scores*, which represent the number of standard deviations away from the mean that a particular datum lies assuming the data are normally distributed. For normal distributions, the data are plotted against the z-scores; for lognormal distributions, the data are log-transformed and plotted against the z-scores. In both cases, parameters of the distribution can be estimated from the least-squares regression line. When the hypothesized distribution is a poor fit to the data, p-plots can yield misleadingly low estimates of the standard deviation (Cullen and Frey, 1999). Both Gilbert (1987) and Ott (1995) provide excellent descriptions of the use of probability plotting to derive parameter estimates for a given distribution. Probability plotting techniques with best-fit lines have been used to estimate parameters for a wide variety of distributions, including beta, Weibull, and gamma.

Cullen and Frey (1999) point out that probability plotting may not be a primary choice for selecting a fitting distributions because the method violates an important assumption of least squares regression—independence of the observations (see Appendix A, Exhibit A-5). This is because the rank-ordered data are no longer independent. Nevertheless, this approach may yield good results when the fit is good and the choice of distributions is somewhat subjective.

B.5.4 PARAMETER ESTIMATION METHODS

As a rule, there are often a number of different methods available for estimating a given parameter. The most appropriate method to apply may require judgment, depending on the relative difficulty in applying a method for a particular parameter, as well as the desired statistical properties of the method. The following simple example provides a useful analogy. Suppose that the parameter of interest, A, is the total area of an approximately square exposure unit. If the exposure unit is a perfect square, and the length of one side (L_1) is known, the area would be equal to L_1^2 (i.e., for a square, $A=L_i^2$). Suppose L is unknown, but two independent measurements, X_1 and X_2, are available to estimate the length (see Exhibit B-5). If it is assumed that the random variable, L, has a probability distribution with mean μ, then the area of the square piece of property is $A=\mu^2$. What is a reasonable estimate of the area (i.e., $\hat{A} = \hat{\mu}^2$) based on X_1 and X_2? Three plausible methods for calculating $\hat{\mu}^2$ are given below.

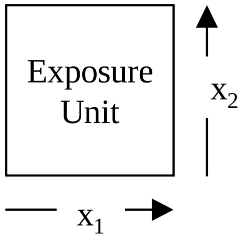
1. $\hat{\mu}_a^2 = \left(\frac{X_1 + X_2}{2}\right)^2$

2. $\hat{\mu}_b^2 = \frac{X_1^2 + X_2^2}{2}$

3. $\mu_c^2 = X_1 \times X_2$

EXHIBIT B-5

ESTIMATING THE AREA OF A HYPOTHETICAL EXPOSURE UNIT



Because these three estimators will, as a rule, give different answers, it may be useful to set criteria for selecting which one gives the "best" answer. Some of the statistical criteria that are used for this purpose are *consistency*, *efficiency*, *robustness*, *sufficiency*, and *unbiasedness* (see Exhibit B-6). It turns out, each method is relatively easy to implement, but the third method is preferred because it is a more efficient estimator.

In many cases, particularly if a model is complex, potential estimators of the unknown parameters are not readily apparent. To assist in developing estimators, several general methods have been developed. Exhibit B-7 lists some of the more common parameter estimation methods.

Perhaps the simplest method is the method of matching moments (MoMM), also called the method of moments. MoMM is appropriately named, as it involves expressing the unknown parameters in terms of population moments and then "matching", or equating the sample moments to the population moments. For example, the sample mean ($\bar{x}$) and standard deviation (*s*) are estimators for the corresponding population parameters (μ and σ).

Maximum Likelihood Estimation (MLE) is a commonly applied method, that is often thought of as a parameter estimate for which the observed data are most "likely". The likelihood function is defined for independent continuous random variables as follows:

$$L(\theta_1, \theta_2,...\theta_k) = \prod_{I=1}^{n} f(x_1|\theta_1, \theta_2, ..., \theta_k)$$

The likelihood function is evaluated based on the product of the PDF for each value of x. The parameters of the probability model, (θ_k), are chosen to maximize the likelihood function value and thereby are most likely to produce the sample data set (Cullen and Frey, 1999).

EXHIBIT B-6

CRITERIA FOR EVALUATING PARAMETER ESTIMATION METHODS*

Consistency A consistent estimator converges to the "true" value of the parameter as the number of samples increases.

Efficiency An efficient estimator has minimal variance in the sampling distribution of the estimate.

Robustness A robust estimator is one that works well even if there are departures from the assumed underlying distribution.

Sufficiency A sufficient estimator is one that makes maximum use of information contained in a data set.

Unbiasedness An unbiased estimator yields an average value of the parameter estimate that is equal to that of the population value.

*Source: Cullen and Frey, 1999

It has also been demonstrated that MLE yields estimators that generally have good properties when evaluated by the criteria listed above. In some cases (e.g., for smaller sample sizes), these estimators are not *unbiased*; however, this can often be accounted for by "adjusting" the estimator. A familiar example of this adjustment is in estimation of the variance of a normal distribution. The MLE for the variance is biased by a factor of ((*n*-1)/*n*), but this is easily corrected by multiplying the MLE by (*n*/(–1)). For some distributions, calculations of the MLE are straightforward. For example, MLE for parameters of a normal distribution are given by the mean and standard deviation of the sample data, the same as MoMM. MLE for parameters of a lognormal distribution are given by the mean and standard deviation of the log-transformed data, which is different from MoMM. In general, MLE calculations are complex, and commercial software such as *@Risk* and *Crystal Ball*® may be used. A more detailed discussion of the derivation and properties of MoMM and MLE can be found in the statistics literature (e.g., Chapter 5 of Mood and Graybill, 1963; Chapter 9 of Mendenhall and Scheaffer, 1973; Section 6.5 of Law and Kelton, 1991; Section 5.6 of Cullen and Frey, 1999).

EXHIBIT B-7

PARAMETER ESTIMATION METHODS

- Method of Matching Moments
- Maximum Likelihood
- Minimum Chi-Square
- Weighted Least-Squares

B.5.5 DEALING WITH CORRELATIONS AMONG VARIABLES OR PARAMETERS

Correlations between exposure variables or between parameters of the probability distribution may be important components of a probabilistic model. Correlation is a measure of association between two quantitative random variables. Two random variables may either be positively or negatively correlated. A positive correlation exists between two variables if the value of X_1 increases as the value of X_2 increases. For example, higher hand dust lead levels have been associated with higher pediatric blood lead levels (Charney et al., 1980). A negative correlation exists between two variables if the value of X_1 increases as the value of X_2 decreases. For example, studies suggest the ingestion of soil and dust particles increases as particle size decreases (Calabrese et al., 1996).

A first step in identifying correlations is to assess the possible physical and statistical relationships that exist between variables. In an ecological risk assessment (ERA), for example, the largest surf scoter (diving duck) does not consume the least amount of food, nor does the smallest surf scoter consume the greatest amount of food. Random sampling of body weight and ingestion rate as separate parameters, however, allows for these two possibilities. Neglecting a correlation between two variables may restrict (underestimate) the tails of the ecological Hazard Quotient (HQ) for each chemical of concern (COC), which are frequently the areas of the distribution of most interest.

The degree to which correlations affect the output of a risk model depends on: (1) the strength of correlations between the two variables, and (2) the contribution of the correlated variables to overall variance in the output (Cullen and Frey, 1999). Therefore, it is useful to conduct a preliminary sensitivity analysis to assess the impact of alternative correlation assumptions on the model output. If the impact is significant, correlations should be identified and accounted for in the PRA.

There are several approaches to account for dependencies in MCA including: (1) modifying the model to include the correlation; and (2) simulating dependence between variables for sample generation (Cullen and Frey, 1999). Modifying the model is preferred as simulation techniques cannot capture the full complexity between model inputs. However, when this is not possible, dependencies between variables can be simulated and approximated by correlation coefficients and bivariate normal distributions.

Correlation coefficients are a numerical measure of the strength and direction of the relationship between two variables. Sample correlation coefficients measure the linear relationship between variables. However, if two variables are from different probability distributions, it is unlikely that they are linearly related. Consequently, simulation software programs such as *Crystal Ball*® and *@Risk* can be used to calculate and employ the nonparametric statistic, Spearman's Rank Correlation Coefficients (Rho) in simulating correlation between inputs. Rank Correlation Coefficients measure the linear dependence not of the data values themselves, but of the rank value of the data. The ranks indicate relative positions in an ordered series, not the quantitative differences between the positions. The disadvantage of losing information by using the rank values (rather than the actual values) is offset by the ability to correlate random variables from different distribution types (See Appendix A).

Exhibit B-8 gives an example of a straightforward approach to specifying a rank correlation between two input variables in a one-dimensional Monte Carlo analysis (1-D MCA) for variability. A range of correlations is explored as a form of uncertainty analysis on the distribution of intakes given a fish advisory of 7.0 μg/day for a chemical.

EXHIBIT B-8

CORRELATION OF INPUT VARIABLES FOR 1-D MCA OF VARIABILITY

Intake Equation Intake = (CF x IR x FI x EF x ED)/(BW x AT)

Variables	Description and Units	Units	Point Estimate or PDFv
CF	concentration in fish	ug/kg	25
IR	fish ingestion rate	kg/meal	lognormal (0.16, 0.07)[1]
FI	fraction ingestion from source	unitless	1.0
EF	exposure frequency	meals/yr	lognormal (35.5, 25.0)[1]
ED	exposure duration	years	30
BW	body weight	kg	70
AT	averaging time	days	10950

[1] Lognormal PDF parameters: arithmetic mean, standard deviation

- Correlation between IR and EF is suggested by Burger et al. (1999) study of 250 anglers on the Savannah River, South Carolina. Moderate correlation (Kendall's tau=0.17, p=0.04)
- Uncertainty Analysis: 1-D MCA simulations of variability correlating IR and EF using *Crystal Ball® 2000* (5,000 iterations, Latin Hypercube sampling). Spearman rank correlations: 0.10, 0.50, 0.90

Statistics of PDFv for Intake (ug/day) compared to Fish Advisory of 7.0 ug/day

Rank Correlation (r)	0.10	0.50	0.90
Intake Statistics (ug/day)			
mean	1.6	1.8	2.0
50th percentile	1.1	1.1	1.1
95th percentile	4.4	5.4	6.5
97.5th percentile	5.7	7.0	9.0

- For this example, only IR and EF are characterized by PDFs. They contribute approximately equally to the distribution of intakes. Positive rank correlations have little effect on the median (50th percentile) of the output distribution, but tend to widen the tails of the distribution. Increasing the correlation from 0.10 to 0.90 increases the 90th percentile from 4.4 to 6.5 ug/day, and the 97.5th percentile from 5.7 to 9.0 ug/day.
- If the fish advisory is 7.0 ug/day, uncertainty in the correlation coefficient may have important consequences for the risk management decision.

Correlations may also be specified for parameters of a probability distribution. This is an important concept when designing a two-dimensional Monte Carlo analysis (2-D MCA) in which parameters of the same PDFv might be otherwise be described by independent PDFu's. A common approach for correlating two parameters is to specify a bivariate normal distribution (Nelsen, 1986, 1987; Brainard and Burmaster, 1992). A bivariate normal distribution allows for the distribution of one variable to be sampled conditional on the other. This is a special case of a joint distribution in which both x and y are random variables and normally distributed (as the conditional distribution of x or of y is always normal) (Wonnacott and Wonnacott, 1981). Example B-4 further explains bivariate normal distributions and demonstrates this approach applied to coefficients of a simple linear regression model that relates contaminant concentrations in soil and dust.

The results of correlation analysis should be interpreted with caution. Two variables may be associated due to: (1) a dependency between the two variables; (2) chance (two independent variables appear dependent due to chance in the sampling procedure); and (3) variables not included in the analysis (lurking variables) are affecting the two variables being analyzed. Likewise, a low correlation measure does not necessarily mean the two variables are independent. As a lurking variable may cause the appearance of an association between the two independent variables, it may also mask the association between two dependent variables.

☞ *Correlation describes a degree of mathematical association, not a causal relationship between the two variables.*

Efforts to extrapolate or predict correlations outside the range of observed values should also be done with caution. For example, there may be a strong linear relationship between age and height in children; however, it would be inappropriate to apply this correlation to adults. Additional caution is needed when correlating more than two factors at a time. In general, because of the complexity of specifying a valid covariance matrix when correlating more than two factors at a time, risk assessors may need to consult a statistician to avoid generating misleading risk estimates.

B.5.6 CENSORED DATA

In order to define the exposure point concentration, estimates of summary statistics representative of the entire distribution of data are needed (Helsel and Hirsch, 1992). Censored data complicate the process of selecting and fitting PDFs and estimating parameter estimates. A censored data set is a data set for which measurements above or below a certain threshold are not available. Left censored data occurs frequently at Superfund sites, where samples for a number of chemicals are often below the reporting limit. A censored datum (often denoted by ND) commonly represents a value of half of the laboratory reporting limit.

Three general methods for estimating summary statistics for left censored data sets include: (1) simple substitution; (2) distributional methods; and (3) robust methods (Helsel and Hirsch, 1992). These methods may be evaluated based on the root mean squared error (RMSE) estimate, a measure of the difference between the sample statistic (e.g., the sample mean, $\bar{x}$) and the true population parameter (e.g., population mean, μ).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \frac{(\overline{x} - \mu)^2}{\mu}}{N}}$$

Methods which yield estimates closer to the true parameter value have lower bias, higher precision, and lower RMSEs.

Simple Substitution Methods

Simple substitution methods entail substituting values equal to or lower than the reporting limit in the data set. These surrogate values are then included in the calculation of the summary statistics and in determining the distributional shape of the data set. Although this method is frequently used, it is important to understand its limitations; depending on the surrogate value used (e.g., half the reporting limit) the simple substitution method may yield biased parameter estimates (e.g., low estimates of the mean) and may yield misleading distributional shapes. Studies such as those reported by Gilliom and Helsel (1986) have determined, in terms of the RMSE, that simple substitution methods perform more poorly than the distributional and robust methods described below.

Distributional Methods

With distributional methods, the entire data set is assumed to follow a theoretical distribution (e.g., normal distribution). Assuming a theoretical distribution, MLE and probability plotting (p-plot) methods provide summary statistics that best match the reported values of the data and the percentage of samples below the threshold value. If the data fit the theoretical distribution exactly, or if the sample size is large, both MLE and p-plots are unbiased methods. Often, however, the sample size is small and the distribution deviates from a theoretical distribution. In this case, the MLE and p-plot methods may yield biased and imprecise methods (Hesel and Hirsch, 1992).

Robust Methods

With robust methods, a theoretical distribution is needed. A theoretical distribution is fit to the data above the detection limit by MLE or p-plot methods. Based on this assumed PDF, the value of the data points below the detection limit are extrapolated and used in the summary statistics calculation. Unlike the simple substitution method, these extrapolated values are not estimates for the data points; rather, they are only used jointly to calculate summary statistics (Hesel and Hirsch, 1992). The method is considered robust as it uses the actual values of the sample data, rather than the distribution above the detection limit.

B.5.7 TRUNCATION

Truncation refers to imposing a minimum and/or maximum value on a probability distribution. The main purpose of truncation is to constrain the sample space to a set of "plausible values". For example, a probability distribution for adult body weight might be truncated at a minimum value of 30 kg and a maximum value of 180 kg in order to avoid the occasional selection of an unlikely value (e.g., 5 or 500 kg). Given the subjectiveness involved in selecting truncation limits, such choices should clearly be made with caution, and involvement of stakeholders who may be aware of site-specific circumstances. For example, there may well be individuals who weigh more than 180 kg and less than 30 kg. The purpose for truncating the tails of a distribution is to confine each risk estimate of a Monte Carlo simulation to a combination of plausible input values. The advantage of truncating unbounded probability distributions in PRA is that central tendency and high-end risk estimates will not be biased by unrealistic values. The disadvantage is that the original parameter estimates of the nontruncated distribution are altered by constraining the sample space. The bias in the parameter estimates increases as the interval between the minimum and maximum truncation limit is reduced. For example, a normal distribution with an arithmetic mean of 100 may be fit to a data set; imposing a truncation limit of 300 may result in a truncated normal distribution with an arithmetic mean of 85. The relationship between the truncated and nontruncated parameter estimates can be determined analytically (Johnson et al., 1995) or approximated using Monte Carlo simulations under both truncated and nontruncated scenarios.

Table B-3. Theoretical bounds and parameter values for selected distributions.

Probability Distribution	Parameters[1]	Theoretical Bounds
Normal	(μ, σ)	$(-\infty, +\infty)$
Lognormal	(μ, σ)	$[0, +\infty)$
Weibull	(α, β)	$[0, +\infty)$
Exponential	(β)	$[0, +\infty)$
Gamma	(α, β)	$[0, +\infty)$
Beta	$(\alpha_1, \alpha_2, a, b)$	[a, b]
Uniform	(a, b)	[a, b]
Triangular	(a, m, b)	[a, b]
Empirical (bounded EDF)	(a, b, {x}, {p})	[a, b]

[1] a=minimum, b=maximum, μ=mean, σ=standard deviation, m=mode, α=shape parameter, β=scale parameter, x=value, p=probability

Truncation is typically considered when using unbounded probability distributions (e.g., normal, lognormal, gamma, Weibull) to characterize variability. Table B-3 gives the theoretical bounds for selected probability distributions that may be more commonly used in PRA. Truncating the minimum value may also be appropriate for distributions whose minimum is defined as zero (e.g., lognormal, gamma, Weibull). Truncation is generally less important when a PDF is used to characterize uncertainty in a parameter estimate (e.g., arithmetic mean), since distributions for uncertainty are often bounded by definition (e.g., triangular, uniform). Bounded continuous distributions, such as the beta distribution or empirical distribution (see Section B.5.2) are not subject to the parameter bias of truncation, although plausible minimum and maximum values must still be identified.

Identifying appropriate truncation limits that reflect "plausible bounds" for an exposure variable will often require judgment. Given that most data sets represent statistical samples of the target population, it is unlikely that the minimum and maximum observed values represent the true minimum and maximum values for the population. However, there may be physiological or physical factors that can aid in setting plausible truncation limits. For example, the maximum bioavailability of chemicals in the gastrointestinal (GI) tract is 100%. Similarly, the solubility of chemicals in aquatic environments (accounting for effects of temperature) will generally be less than the chemical solubility in water free of particulates.

In general, sensitivity analysis can be used to determine if truncation limits are an important source of parameter uncertainty in risk estimates. For exposure variables in the numerator of the risk equation, the maximum truncation limit is of greatest concern. For exposure variables in the denominator of the risk equation, the minimum truncation limit is of greatest concern. Details regarding the fit of the tails of the probability distribution and the effect of truncation on the parameter estimates should generally be included in the workplan.

B.6.0 ASSESSING QUALITY OF THE FIT

The quality of the fit of a distribution may be evaluated in several ways. Standard statistical approaches are available to test the fit of a theoretical distribution to a data set (i.e., GoF tests). In addition, alternative choices for distribution shapes and plausible bounds might be explored as a form of sensitivity analysis. Together with graphical exploration (Section B.5.3), this information may be useful when deciding whether or not to incorporate a specific type of distribution for an exposure variable into a PRA.

☞ *GoF tests are one tool among several to assess the quality of a distribution.*

Although GoF testing is a necessary part of distribution fitting, and tests are readily available with commercial software, it is less important than mechanistic considerations or graphical data exploration for choosing a candidate distribution. Examples of GoF tests are discussed below, and cautions regarding GoF are outlined in Section B.6.3.

B.6.1 WHAT IS A GOODNESS-OF-FIT TEST?

Goodness-of-fit (GoF) tests are formal statistical tests of the hypothesis that the data represent an independent sample from an assumed distribution. These tests involve a comparison between the actual data and the theoretical distribution under consideration.

In statistical hypothesis testing the null hypothesis (H_0) is assumed to be true unless it can be proven otherwise. The "evidence" upon which we base a decision to reject or not to reject H_0 is a random sample. Typically, we seek to reject H_0 in favor of H_a. For example, with the two sample *t*-test, the null hypothesis is that the means of two populations are equal (not different) and the alternative is that they are different. This is expressed as:

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

Most often, the hypothesis test is used to show that the means are not equal (i.e., reject H_0 in favor of H_a) in order to state that there is a significant difference between the two populations at a specified significance level (e.g., α=0.05). Thus, the hypothesis test is often referred to as a significance test.

The *p-value* in a statistical test is calculated from a sample and represents the probability of obtaining a value of the test statistic as extreme or more extreme as the one observed if H_0 is in fact true. When the *p-value* is small it means either the null hypothesis is not true, or that we have witnessed an unusual or rare event (by chance we drew an unusual sample that resulted in the extreme value of the test statistic). Often a value of 0.05 or 0.01 is designated as a cutoff, or significance level α. If the *p-value* is (e.g., $p < 0.05$), the null hypothesis is rejected in favor of the alternative, and we state that the test result is statistically significant at level α. This does not mean that we have proven H_a is true. Rather, we are saying that based on our sample results, it is unlikely that H_0 is true.

In a GoF test, the hypothesis test is set up the same way as a “traditional” hypothesis test, but the outcome is viewed a little differently. In GoF tests, we generally seek to *fail* to reject H_0 because the null hypothesis states that the data were obtained from a population described by the specified distribution (F_0). The alternative hypothesis is that the data were obtained from a population described by a different distribution. In most applications of GoF techniques, the alternative hypothesis is composite—it gives little or no information on the distribution of the data, and simply states that H_0 is false (d’Agostino and Stephens, 1986). This can be expressed as:

$$H_0 : F = F_0$$
$$H_a : F \neq F_0$$

where F_0 is a specific continuous distribution function, such as the CDF for a normal distribution.

☞ *GoF tests do not prove that the population is described by the specified distribution, but rather that this assumption could not be rejected.*

In general, *p-values* provide one metric of evaluating the fit of the distribution. For example, a *p-value* of 0.06 indicates that the null hypothesis (i.e., the assumption of a specified distribution) cannot be rejected at α=0.05. Larger *p-values* indicate a better fit and stronger evidence that the distribution specified by the null hypothesis may be appropriate. This guidance does not recommend an arbitrary cutoff for the *p-value*. A risk assessor performing a GoF test generally should report the *p-value* and whether the fit is considered “good” or “poor”.

B.6.2 WHAT ARE SOME COMMON GOODNESS-OF-FIT TECHNIQUES?

The following GoF tests can also be found in most general statistical and spreadsheet software. Both *Crystal Ball*® and *@Risk* software present the results of chi-square, K-S, and Anderson-Darling tests in their fitting routines.

Shapiro-Wilk Test

The most widely used GoF test in risk assessment is the Shapiro-Wilk test for normality (Gilbert, 1987). This simple hypothesis test can determine whether or not a small data set ($n \leq 50$) is normally distributed. The test can also be run on log-transformed data to assess whether the data are lognormally distributed. D'Agostino's test may be used for samples sizes larger than those accommodated by the Shapiro-Wilk test (i.e., $n > 50$) (d'Agostino and Stephens, 1986). In addition, Royston (1982) developed an extension of the Shapiro-Wilk test for *n* as large as 2000 (Gilbert, 1987).

Probability Plot Correlation Coefficient Test

The correlation coefficient *r* (or the coefficient of determination, r^2) between the data and the z-scores of a normal probability plot (Filliben, 1975; Helsel and Hirsch, 1992) is similar to the *W* statistic of the Shapiro-Wilk test. A detailed comparison of the Shapiro-Wilk test and the product correlation coefficient test is given by Filliben (1975) and d'Agostino and Stephens (1986). Helsel and Hirsch (1992) summarize critical r* values derived by Looney and Gulledge (1985) for the probability plot correlation coefficient test.

Chi-Square Test

The chi-square test is a general test that may be used to test any distribution (continuous or discrete), and for data that are ordinal (e.g., categories such as high/medium/low). Chi-square is a measure of the normalized difference between the square of the observed and expected frequencies. For example, by constructing a frequency distribution of the data with *k* adjacent bins, j=1...k, the number of data points in the j^{th} bin can be compared with the expected number of data points according to the hypothesized distribution. Note that in the case of continuous, unbounded distributions (e.g., normal), the first and last intervals may include $[-\infty, a_1]$ or $[a_k, +\infty]$ (Law and Kelton, 1991). The chi-square test is very sensitive to the chosen number and interval width of bins—different conclusions can be reached depending on how the intervals are specified. Strategies for selecting bins (e.g., setting interval widths such that there are no fewer than 5 data points expected per bin) are given in the statistical literature (d'Agostino and Stephens, 1986; Law and Kelton, 1991). The test statistic is compared with a value of the chi-square distribution with ($k - r - 1$) degrees of freedom, where *k* is the number of sample values and *r* is the number of parameters of the hypothesized distribution. As described in Section B.6.1, in general, higher *p-values* suggest better fits.

Kolmogorov-Smirnov (K-S) Test

The K-S test is a nonparametric test that compares the maximum absolute difference between the step-wise empirical CDF and the theoretical CDF. Because the maximum discrepancy is compared with the test statistic, K-S is sometimes referred to as a *supremum* test (Cullen and Frey, 1999). In general, lower values of the test statistic indicate a closer fit. The K-S test is most sensitive around the median of a distribution, and, hence, it is of little use for regulatory purposes when the tails of distributions are most generally of concern (U. S. EPA, 1999a). Although it does not require grouping data into bins like the chi-square test, critical values for the K-S test depend on whether or not the parameters of the hypothesized distribution are estimated from the data set (Gilbert, 1987; Law and Kelton, 1991). The Lilliefors test was developed to surmount this problem when the hypothesized distribution is normal or lognormal (Gilbert, 1987).

Anderson Darling Test

The Anderson-Darling test assesses GoF in the tails (rather than the mid-ranges) of a PDF using a weighted average of the squared differences between the observed cumulative densities. The Anderson-Darling test is sometimes referred to as the *quadratic* test (Cullen and Frey, 1999). The test statistic should be modified based on sample size prior to comparison with the critical value. Like the K-S test, in general, lower values of the test statistic indicate a closer fit (i.e., if the adjusted test statistic is greater than the modified critical value for a specified α, the hypothesized distribution is rejected). The Anderson-Darling test may be particularly useful because it places more emphasis on fitting the tails of the distribution.

B.6.3 CAUTIONS REGARDING GOODNESS-OF-FIT TESTS

There are many statistical software programs that will run GoF tests against a long list of candidate distributions. It is tempting to use the computer to make the choice of distribution based on a test statistic. However, GoF tests have low statistical power and often provide acceptable fits to multiple distributions. Thus, GoF tests are better used for rejecting poorly fitting distributions than for ranking good fits. In addition, for many distributions, GoF statistics lack critical values when the parameters are unknown (i.e., estimated from the data). In practice, this limitation is often discounted and the critical values are interpreted as a semi-quantitative measure of the fit. It is most appropriate to form an idea of the candidate distributions based on some well reasoned assumptions about the nature of the process that led to the distribution, and then to apply a GoF test to ascertain the fit (U.S. EPA, 1999a). Whenever possible, mechanistic and process (i.e., phenomenologic) considerations should inform the risk assessor's choice of a particular distribution rather than the results of a comparison of GoF tests (Ott, 1995). In addition, the value of graphical evaluations of the fit cannot be overstated.

B.6.4 ACCURACY OF THE TAILS OF THE DISTRIBUTION

The tails of a distribution (e.g., < 5th and > 95th percentiles) for an input variable are often of greatest interest when characterizing variability in risk. Distributions fit to data may not characterize the tails of the distribution in a way that represents the target population. In general, the importance of uncertainty in the fit of the tails of particular distributions should be determined on a site-specific basis. For exposure variables in the numerator of the risk equation, the upper tail is of greatest concern. For exposure variables in the denominator of the risk equation, the lower tail is of greatest concern.

The tails of the input PDFs generally have a significant influence on the tails of the risk distribution, especially for those variables that are ranked highest in a sensitivity analysis. Different distributions may share the same mean and variance, but assume very different shapes. Experiments with Monte Carlo simulations have demonstrated that the shape of the input PDFs may have a minimal effect on the risk estimates in the tails of the probability distribution when the mean and variance of the input PDFs are held constant (Hoffman and Hammonds, 1992; Finley and Paustenbach, 1994). Nevertheless, it is generally a good practice in PRA to demonstrate that alternative choices of PDFs do not have a significant effect on percentiles in the RME risk range.

A common question when developing and evaluating Monte Carlo models is, "How many iterations is enough?". Since Monte Carlo sampling is approximately random, no two simulations will yield the same results (unless the same starting point, or seed, of the random number generator is used). A rule of thumb is that the stability of the output distribution improves with increasing numbers of iterations, although there will always remain some stochastic variability. The stability is generally better at the central tendency region of the output distribution than at the tails; therefore, more iterations may be needed when the risk management decision is associated with the higher percentiles (e.g., > 95th percentile). Risk assessors are encouraged to run multiple simulations (with the same inputs) using different numbers of iterations in order to evaluate the stability of the risk estimate of concern. The results of such an exercise should generally be reported to the Agency when submitting a PRA for review. Note that while the speed of modern computers has essentially eliminated the issue for 1-D MCA (e.g., 10,000 iterations of most 1-D MCA models can be run in less than 1 minute), it may still be an important issue for more complex modeling approaches such as Microexposure Event analysis (MEE) and 2-D MCA (see Appendix D).

B.7.0 SELECTING PROBABILITY DISTRIBUTIONS BASED ON STATE OF KNOWLEDGE

Table B-4 summarizes preliminary strategies for proceeding with a PRA based on the amount of available information. Recommended starting points for each of the three steps in the general process are provided. This table provides guidance on candidate distributions that are consistent with the available information, however, it is not intended to discourage the use or exploration of alternative choices.

☞ *Table B-4 provides recommended preliminary strategies, not steadfast rules. As an analyst works through the PRA, alternative distributions, estimation methods, consideration of mechanism, and GoF tests may better guide the selection process.*

Case 1 represents the best scenario, in which the analyst has access to the raw data and a sufficiently large sample size (or ≥ 6 percentiles). In this case, the analyst has a variety of choices for distribution fitting and estimating parameters. However, frequently raw data are inaccessible to the analyst. Cases 2 and 3 have limited information available (i.e., mean and upper percentile) and, therefore, have a narrower set of starting points. Case 4 is the most extreme scenario of data availability requiring expert judgment on selecting and fitting distributions.

Table B-4. Strategies for conducting PRA based on available information. Preferred methods in Case 1 (most information) are identified by an asterisk (*).

Evaluation Step	Case 1	Case 2	Case 3	Case 4
	Decreasing Information →			
Data Availability	raw data of sufficiently large sample size *or* six or more percentiles	three to five statistics	two statistics	one statistic
Selection of Distribution Type	**Nonnegative Continuous** any in this category **Bounded** beta, Johnson's SB	**Nonnegative Continuous** lognormal, gamma, Weibull **Bounded** beta, Johnson's SB		case-by-case basis using expert judgment
Selection of Parameter Estimation / Fitting Method	maximum likelihood* regression methods matching moments	minimize average absolute percent error (MAAPE) for available statistics	exact agreement between 2-parameter PDF and available statistics	
Assessment of Quality of Fit	**Graphical Assessment** P-log Q plot*, P-Q plot* residual % error plot* P-P plot, Q-Q plot **GoF Tests** Anderson-Darling* K-S Chi-square	**Graphical Assessment** P-log Q plot, P-Q plot **GoF Test** Chi-square, Estimate *p*-value for MAAPE using parametric bootstrap (if sample size is known)	**Graphical Assessment** judgment based on comparative analysis of PDFs and CDFs	
Estimation of Parameter Uncertainty	**Large Sample** asymptotic normality assumption **Medium Sample** nonparametric bootstrap **Small Sample** parametric bootstrap	**Parametric bootstrap** generate random samples using the fitted distribution (if sample size is known)		

EXAMPLES OF FITTING DISTRIBUTIONS USING GRAPHICAL METHODS, GOODNESS-OF-FIT, AND PARAMETER ESTIMATION

Example B-1. Empirical Distribution Function (EDF) for Soil Ingestion Rates

This hypothetical example illustrates how graphical methods can be used to select probability distributions for variability based on percentile data reported in the literature. Table B-5 gives the summary statistics that are reported by Stanek and Calabrese (1995) for average daily soil ingestion rates among young children. Three options are explored for selecting a distribution: (1) empirical distribution function (EDF) represented by a step function; (2) linearized and extended EDF; and (3) continuous parametric distributions (beta and lognormal).

In order to specify an EDF, a plausible range (minimum and maximum) must be inferred using judgment. Exposure factors such as ingestion rate are nonnegative variables (i.e., minimum $\geq$0); given the relatively low value for the 25th percentile (10 mg/day), it is assumed that 0 mg/day is a reasonable minimum value for this example. If children with pica for soil are excluded from the population of concern, the maximum value may be inferred from the relatively shallow slope at the high-end of the distribution. That is, the 90th percentile is reported as 186 mg/day while the 99th percentile is 225 mg/day, an increase of only 39 mg/day; it is assumed that 300 mg/day is a plausible maximum value for this example. Commercial software such as *Crystal Ball*® and *@Risk* can be used to input EDFs. Figure B-3 illustrates the basic step-wise EDF represented by the reported percentile values, as well as the "linearized, extended EDF" (i.e., linear interpolation between reported values and extended lower and upper tails).

An alternative to relying on a linear interpolation between the percentile values is to fit a continuous probability distribution to the reported percentiles. Since the original data are unavailable, standard GoF tests for the EDF, such as K-S and Anderson-Darling (d'Agostino and Stephens, 1986), cannot be applied. Note that computer software (e.g., *Crystal Ball*®, *@Risk*) will provide test statistics and corresponding *p-values*, however, these results will (inappropriately) reflect the number of percentile values reported rather than the sample size of the original data. Nevertheless, graphical methods may be employed to assess the adequacy of the fit of various PDFs. In this example, a beta distribution and lognormal distribution were fit to the EDF using *Crystal Ball*®. Figure B-4 illustrates the selected statistics for both distributions.

The beta distribution appears to more closely match the reported percentile values, especially at the upper tail of the distribution. The lognormal distribution has an unbounded maximum that, for this example, results in an extreme overestimate of the 95th and 99th percentiles. The beta distribution, by definition, is bounded at 0 and 1, and rescaled in this example to a maximum of 364 mg/day. This analysis would support the use of a beta distribution in a Monte Carlo simulation.

Table B-5. Selected statistics for reported and fitted distributions for ingestion rate (mg/day).

Summary Statistic	Reported Values	Linearized, Extended EDF	Beta Distribution[1]	Lognormal Distribution[2]
minimum	--	0	0	0
25th percentile	10	10	13	11
50th percentile	45	45	44	31
75th percentile	88	88	100	86
90th percentile	186	186	165	216
95th percentile	208	208	205	375
99th percentile	225	225	322	3346
maximum	--	300	364	+ ∞

[1] Parameters of best-fit beta distribution: α_1=0.63, α_2=2.85, min=0, max=364.
[2] Parameters of best-fit lognormal distribution: μ=97.6, σ=291.8.

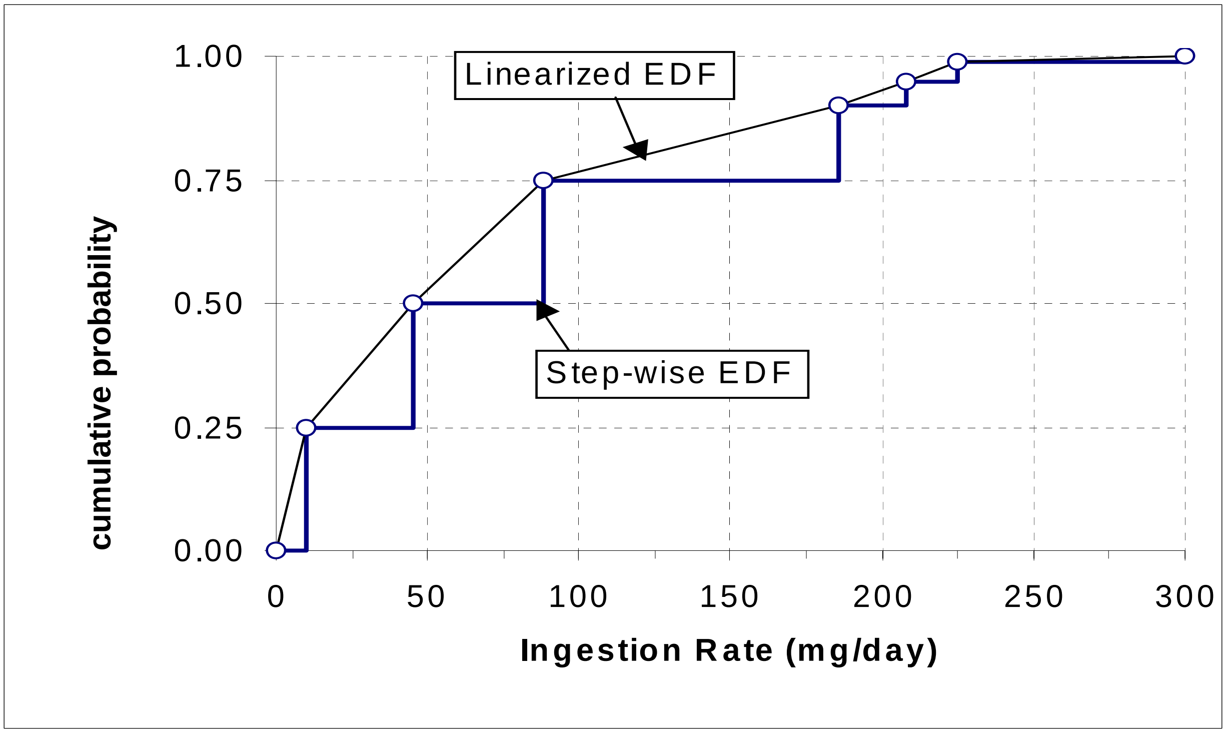


Figure B-3. Comparison of step-wise EDF and linearized EDF for ingestion rate. The upper and lower tails of both distributions are extended to a plausible range of [0, 300] mg/day.

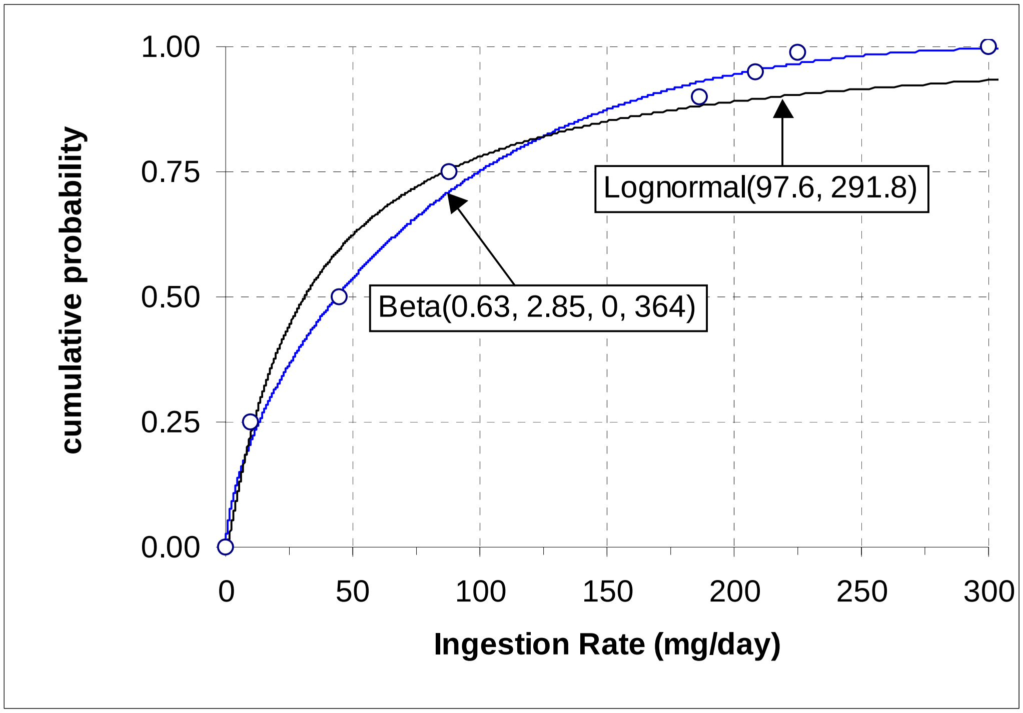


Figure B-4. Graphical assessment of beta and lognormal distributions fit to the cumulative distribution reported in the literature (circles). The beta distribution provides a closer fit to the percentile values in this example.

Example B-2. Variability in Lead Concentrations in Quail Breast Tissue

This hypothetical example demonstrates how the combination of graphical methods, GoF tests, and parameter estimation techniques provides strong evidence for selecting and fitting a lognormal distribution. Assume lead concentration in quail is an important variable for a food web model. Site-specific data (*n*=62) are used to estimate inter-individual variability in concentration (Table B-6). The histograms in Figure B-5 show lead concentrations in quail breast tissue collected near a settling pond at a plating works. Equation B-1 indicated that 7 bins is an appropriate starting point. The result (top left panel, Figure B-5) suggests that approximately 80% of the values are < 200 ppm and that the probability distribution for variability may be described by a nonnegative, right-skewed distribution (e.g., exponential, Weibull, lognormal, etc.). However, additional bins are needed to better understand the low-end of the distribution. After increasing the number of bins from 7 to 16 (top right panel, Figure B-5), graphical evaluation continues to suggest that the distribution is unimodal right skewed. The bottom panel of Figure B-5 illustrates that increasing the number of bins would not provide better resolution of the low-end of the distribution. For these data, 16 bins appear to provide a reasonable balance between too much smoothing and too much jaggedness.

Probability plots can be used to visually inspect the GoF of a specified distribution to the data, and, because the hypothesized distribution yields a straight line, the plots are particularly useful for evaluating deviations at the tails. In addition, parameter estimates can be obtained from the regression lines fit to the data, as discussed below. For this example, two lognormal probability plots are explored to evaluate how well the data can be described by a lognormal distribution (Figure B-6). The top panel gives the *z-score* on the abscissa (the “x” axis) and ln[concentration] on the ordinate (the “y” axis), while the bottom panel gives ln[concentration] on the abscissa and *z-score* on the ordinate. Plotting positions for both methods were calculated using Equation B-2. Equally plausible parameter estimates can be obtained from regression lines using either plotting method; however, the approach shown in the top panel may be easier to implement and interpret.

Despite the relatively large sample size of *n*=62, GoF tests generally fail to reject lognormality (i.e., normality of the log-transformed data) in this example. For the probability plot correlation coefficient test (Filliben, 1975; Looney and Gulledge, 1985), if $r < r^*$ (the value for r at a specified α), normality is rejected. For this example, r is 0.988, and r^* is between 0.988 and 0.989 for n=62 and α=0.25; therefore, the *p-value* for the concentrations is approximately 0.25 and one fails to reject lognormality at $\alpha \le 0.25$. D’Agostino’s test yields essentially the same conclusion, with a calculated Y value of -1.9166. For this data set, with n=62 and α=0.10, one rejects normality if $Y < -2.17$ or $Y > 0.997$ (see Table 9.7 in d’Agostino and Stephens, 1986); therefore, since Y is within this interval, one fails to reject the normal distribution. However, for α=0.20, the rejection criteria is [$Y < -1.64$ or $Y > 0.812$], Y falls outside the low-end of the interval, resulting in a rejection of the normal distribution. For this data set, the *p-value* associated with d’Agostino’s test is slightly less than 0.20 and one fails to reject normality at $\alpha < 0.20$.

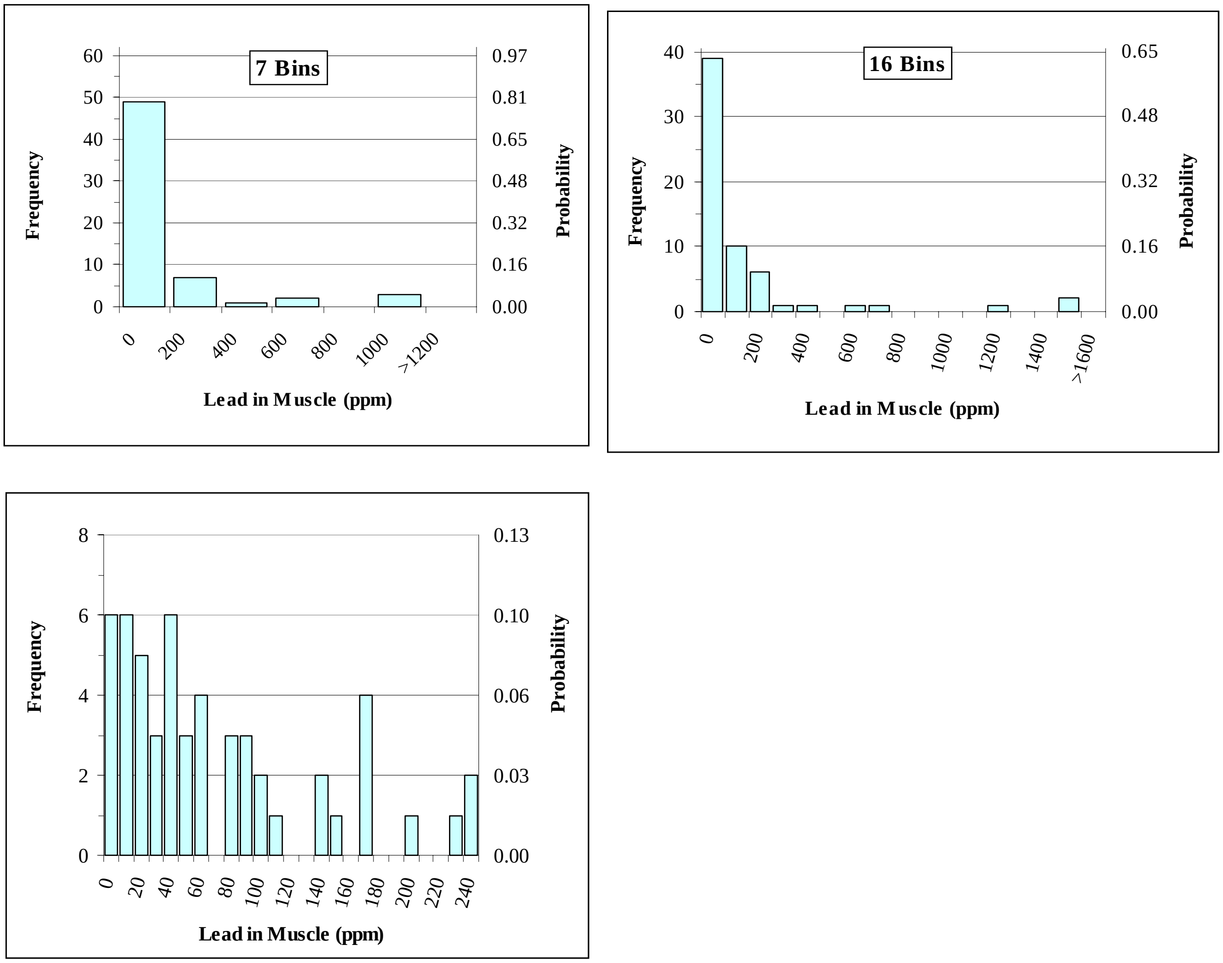


Figure B-5. Histograms of lead concentrations in quail breast muscle (*n*=62). The top left panel shows the result with seven bins; the top right panel shows the result with sixteen bins; the bottom panel uses bin widths of 10 ppm to highlight the lower tail (< 250 ppm) of the distribution.

Table B-6. Sample values of lead concentration (ppm) in quail breast muscle (*n*=62).

0.45	15.8	36.6	57	91	173	265
2.1	16	40	59.6	94.2	175.6	322
5.4	16.7	40.1	61.4	99	176	490
7.8	21	42.8	62	107	177	663.4
7.8	23	44	64	109	205	703
8.8	24	46	64	111	239	1231
11.8	24.8	47	84.6	149	241	1609
12	29.2	49	86.6	149	245	1634
15	35.5	53	86.8	154	264	

Different methods for obtaining the parameter estimates for the lognormal distribution can be explored in this example. For the lognormal distribution, MLE and MoMM simply require calculating the mean and standard deviation of the log-transformed sample data. For the lognormal probability plot method, the parameters can be obtained directly from the least squares regression line expressed as follows:

$$\ln(x) = [slope]z + [intercept] \qquad \text{Equation B-5}$$

such that exponentiating the intercept will give the geometric mean (GM) and exponentiating the slope will give the geometric standard deviation (GSD) (see Footnote 3 of Table B-7). Both the MLE and MoMM estimates will generally match the arithmetic mean of the log-transformed data (i.e., intercept) determined from lognormal probability plots; however, estimates of the standard deviation (i.e., slope) will vary (Cullen and Frey, 1999). In general, the probability plot method yields estimates of the standard deviation that are less than or equal to that of MoMM and MLE, and the results yield closer estimates as the correlation coefficient of the probability plot increases (Cullen and Frey, 1999). Table B-7 summarizes the parameter estimates using MLE, MoMM, and the two lognormal probability plotting techniques described above. The corresponding parameter estimates for the untransformed data are also presented.

In this example, the strong linearity of the probability plots (r^2=0.98) shown in Figure B-6 is an indication that a lognormal distribution is a reasonable model for describing variability in concentrations. The tails of the distributions fit the data fairly well, although the bottom panel suggests that the lognormal distribution slightly overestimates the lower tail. Furthermore, the parameter estimates of the lognormal distribution using probability plotting closely match the estimates using MLE and MoMM.

Table B-7. Parameter estimates for lognormal distribution of lead concentrations (ppm).

Parameter Estimation Method	Log-transformed Data		Untransformed Data[3]	
	Arithmetic mean [$\hat{\mu}$]	Arithmetic stdev [$\hat{\sigma}$]	Arithmetic mean [$\hat{\mu}$]	Arithmetic stdev [$\hat{\sigma}$]
Maximum Likelihood Estimate (MLE)	4.175	1.522	207	626
Method of Matching Moments (MoMM)	4.175	1.522	207	626
Log Probability Plot[1]	4.175	1.507	203	597
Log Probability Plot[2]	4.175	1.543	214	670

[1] Least squares regression line for Figure B-6, top panel.

[2] Least squares regression line for Figure B-6, bottom panel.

[3] For a lognormal distribution, the following equations can be used to convert parameters of the normal distribution of log-transformed data to corresponding parameters of the lognormal distribution of untransformed data. Assume μ* and σ* are the arithmetic mean and standard deviation, respectively, for the normal distribution of log-transformed data.

$$\textit{geometric mean} = \exp[\mu *]$$

$$\textit{geometric standard deviation} = \exp[\sigma *]$$

$$\textit{arithmetic mean} = \exp[\mu * + 0.5\sigma *^{2}]$$

$$\textit{standard deviation} = \exp[\mu *]\left(\exp[\sigma *^{2}]\exp[\sigma *^{2} - 1]\right)^{0.5}$$

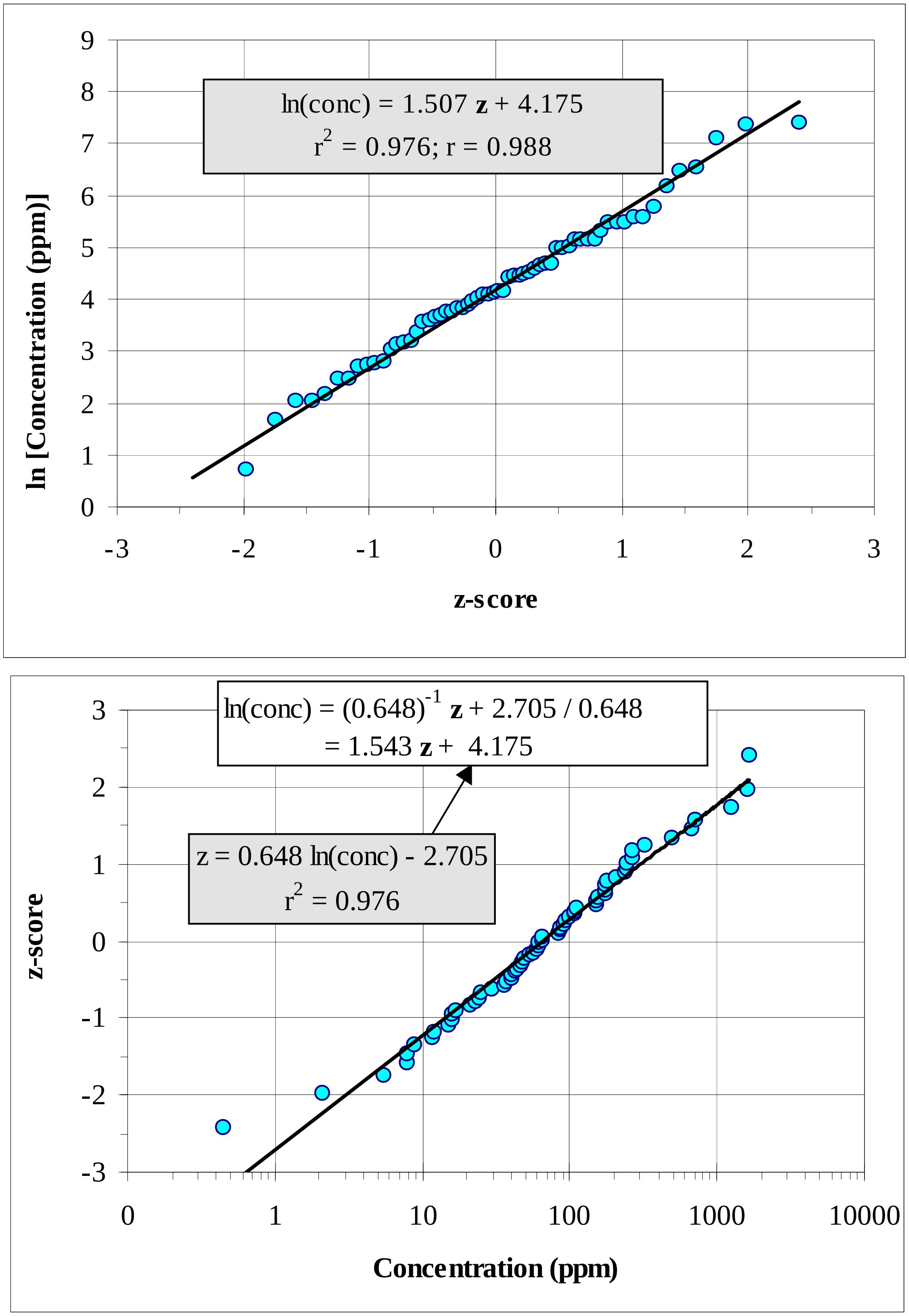


Figure B-6. Lognormal probability plots of lead in quail breast tissue. Top panel gives ***z*** on the abscissa and ln[concentration] on the ordinate. Bottom panel gives concentration (log scale) on the abscissa and ***z*** on the ordinate. Equally plausible parameter estimates can be obtained from regression lines using either plotting method. Bottom panel requires an additional step to express the equation that yields parameter estimates [ln(x)=(slope) z + (y-intercept)], where the slope estimates the standard deviation of ln(x) and the y-intercept (at z=0) estimates the arithmetic mean of ln(x).

Example B-3. Variability in Meal Sizes Among Consuming Anglers

A creel survey of anglers consuming contaminated fish was performed to estimate variability in fish meal sizes. The anglers were asked how many people would eat their fish. The lengths of the fish were measured and a regression equation was used to calculate the corresponding weights. The portion of the fish mass that is consumed was assumed to be 40% (e.g., fillets). Results given in Table B-8 are expressed in units of grams of fish per meal.

The appearance of the histograms (Figure B-7) suggests that the sample (*n*=52) may have been selected from a single distribution.

A normal probability plot of the meal sizes (Figure B-8) shows a departure from linearity. Specifically, there appears to be a "kink" in the probability plot at about 400 g/meal, suggesting that the sample may have been obtained from two unique distributions. Both the Filliben test and Shapiro-Wilk test indicated a significant departure from normality at α=0.01. Parameters may be read directly from the equations of the regression lines on the right hand panel of the graph. MoMM and MLE gave similar estimates.

Table B-8. Meal size (g/meal) (*n*=52).

65	182	310	405
74	208	314	415
74	221	318	416
77	226	318	477
90	241	327	531
110	248	332	572
111	253	336	608
133	260	337	745
143	261	350	831
150	281	351	907
163	303	360	1053
163	305	365	1189
174	305	390	1208

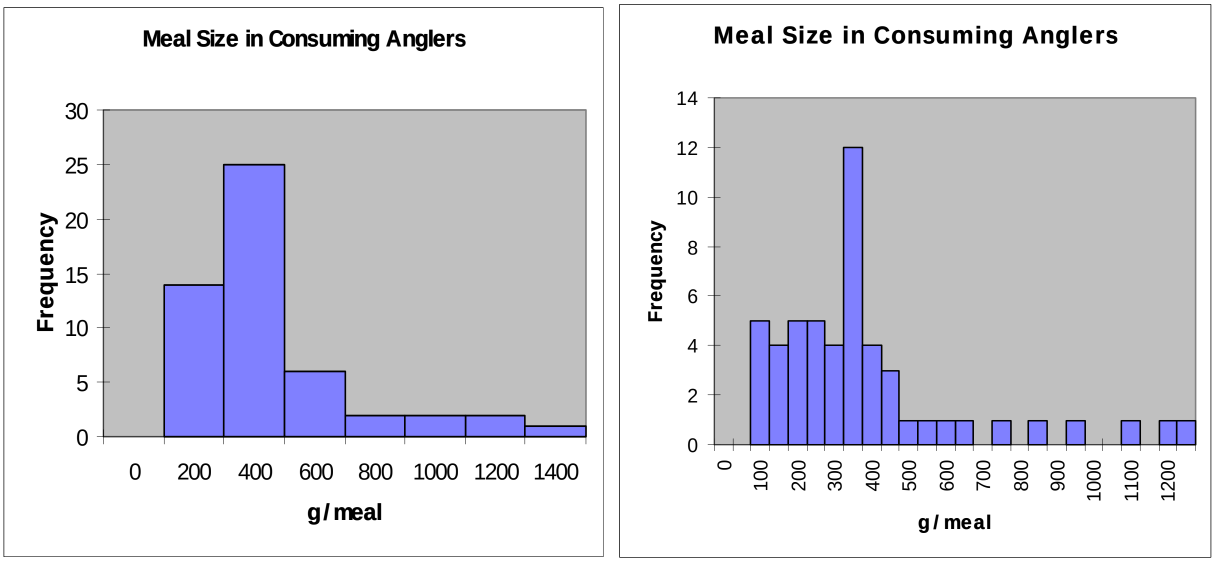


Figure B-7. Histograms of meal size (*n*=52) among consuming anglers. Left panel uses 7 bins, while the right panel uses 14 bins.

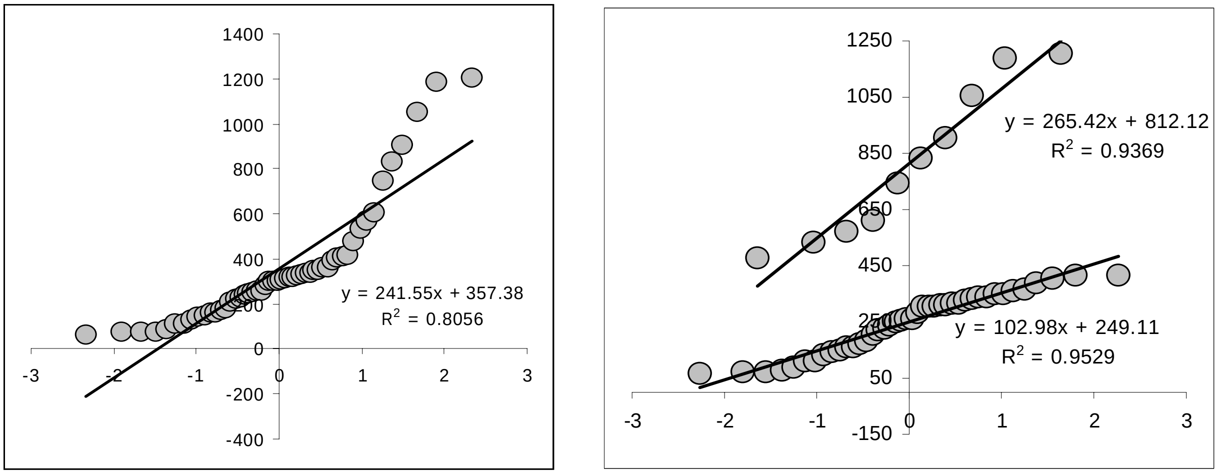


Figure B-8. Probability plot of meal size data from consuming anglers. The left panel shows the combined data, with a departure from linearity at ~ 400 g/meal. The right panel shows the data split between high consumers (top line) and low consumers (bottom line); note that separate lognormal probability plots were reconstructed for both subsets of the data. The point at which to "split" the distribution in the left panel is somewhat subjective. The break would be more obvious if the two distributions did not overlap.

Example B-4. Bivariate Normal Distributions

This example introduces the bivariate normal distribution to illustrate two concepts: (1) use of information on correlations in a Monte Carlo simulation; and (2) specifying distributions for uncertainty in parameter estimates. A brief explanation of the bivariate distribution is presented followed by an example comparing assumptions of no correlation and perfect correlation. A less complex example of a method for addressing correlations in PRA is given in Exhibit B-8.

THIS EXAMPLE PRESENTS...

- Description of the assumptions associated with the bivariate normal distribution
- Guidance on simulating the bivariate normal distribution for two random variables
- Application of bivariate normal to a simple linear regression equation relating contaminant concentrations in soil and dust (see Figure B-9). Results are compared to the assumption of no correlation and perfect correlation

Properties of a Bivariate Normal Distribution

One approach that can be used to correlate two random variables is to specify a bivariate normal distribution, which allows for the distribution of one variable to be sampled conditional on the other. A bivariate normal distribution is a special case of a joint distribution in which both x and y are random independent normally distributed variables. A bivariate normal distribution can be specified for all correlation coefficients including ρ=0, ρ=1, and ρ=−1. The bivariate distribution has a three dimensional shape and for ρ=0, from a bird's-eye view, is perfectly circular. As correlation increases (i.e. moves towards -1 or 1) this circle narrows and flattens to an elliptical shape, and finally for perfect correlation →=1 and ρ=-1) becomes a straight regression line with a r^2=1. In three dimensional space the probability of obtaining measurement pairs (x, y) in the region is equal to the volume under the surface in that region. To completely specify the bivariate normal, estimates of the arithmetic mean and variance of the two parameters, as well as the correlation coefficient (μ_X and μ_Y, variances σ^2_X and σ^2_Y , and correlation coefficient ρ) are needed.

In a bivariate normal distribution, values of y corresponding to each value of x follow a normal distribution (Snedecor and Cochran, 1989). Analogously, the values of x corresponding to each value of y follow a normal distribution. Furthermore, if two random variables, *X* and *Y*, jointly follow a bivariate normal distribution, the marginal distribution of *X* is normal with mean μ_X and variance σ^2_X, and the marginal distribution of *Y* is normal with mean μ_Y and variance σ^2_Y.

Conditional Distributions

Assume we are interested in the conditional distribution of *X* given a certain value for *Y*. For example, if *X* and *Y* are positively correlated, we would expect that relatively high values of *X* tend to correspond with relatively high values of *Y*. The conditional distribution of *X* given that *Y*=y, where y represents a specific value for the random variable *Y*, is a normal distribution with:

$$mean = \mu_X + \rho \frac{\sigma_X}{\sigma_Y}(y - \mu_Y), \quad and$$

$$variance = \sigma_X^2 (1 - \rho^2)$$

Equation B-6

Likewise, the conditional distribution of *Y* given that *X*=x, is also normal with:

$$mean = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X), \quad and$$

$$variance = \sigma_Y^2 (1 - \rho^2)$$

Equation B-7

These general equations can be used to generate a correlated pair (*X*, *Y*), as described below.

*Note that the mean of the conditional distribution of *X* is a function of the given value of *Y* but the variance depends only on the degree of correlation.

General Approach for Correlating X and Y

To generate a correlated pair (*X*, *Y*), first generate *X* using a random value Z_1 from the standard normal distribution:

$$X = \mu_X + \sigma_X \times Z_1$$

Equation B-8

Next, express Y as a function of the conditional mean and variance of *Y* given *X* and a second standard normal variate Z_2:

$$Y = \mu_Y + \sigma_Y \times Z_2$$

Equation B-9

and generate a correlated *Y* by plugging Equation B-7 into Equation B-9. Using algebra, the combined equations yield the following simplified expression for generating *Y*:

$$Y = \mu_Y + \sigma_Y\left[(\rho \times Z_1) + \sqrt{1-\rho^2} \times Z_2\right] \qquad \text{Equation B-10}$$

The important component of this equation is that two random variates are needed (Z_1 and Z_2).

An alternative, but less general approach would be to obtain *Y* by first generating a normal variate *X* (Equation B-8) and then plugging that value into the regression equation of *Y* on *X* to obtain the associated value of *Y*. While this method maintains a correlation between *X* and *Y*, it will underestimate parameter uncertainty. The results are equal only for the special case of perfect correlation (ρ=1.0) between *X* and *Y*. Therefore, the more general bivariate normal distribution approach (given by Equations B-8 to B-10) is recommended for correctly correlating *X* and *Y* because it provides a more robust estimate of parameter uncertainty.

EXHIBIT B-9

STEPS FOR SIMULATING UNCERTAINTY IN LINEAR REGRESSION EQUATION USING A BIVARIATE NORMAL DISTRIBUTION TO CORRELATE PARAMETERS (B_0, B_1)

(1) Select Z_1 from a standard normal distribution $Z \sim N(0, 1)$

(2) Calculate β_0 using Equation B-8, where $X=\beta_0$, $\mu_x=\mu_{b0}$, and $\sigma^2_x=\sigma^2_{b0}$

(3) Select Z_2 from a standard normal distribution $Z \sim N(0, 1)$

(4) Calculate β_1 using Equation B-10, where $Y=\beta_1$, $\mu_y=\mu_{b1}$, $\sigma^2_y=\sigma^2_{b1}$, ρ=correlation between β_0 and β_1

Application of Bivariate Normal Distribution to Correlate Concentrations of Zinc in Soil and Dust

Assume random sampling of soil and dust zinc concentrations occurs in a residential area. Composite samples of soil and dust are collected from 21 locations such that samples are paired (i.e., each soil sample is co-located with a dust sample) (Table B-9). First the relationship between the zinc concentration in soil and dust is evaluated using simple least-squares regression. Next, the bivariate normal distribution for the slope (β_1) and intercept (β_0) is determined, yielding an arithmetic mean and standard deviation for each parameter (μ_{b0}, $\sigma^2_{b0,}$ μ_{b1}, and σ^2_{b1}), and correlation coefficient ρ between β_1 and β_0. In this context, the bivariate normal distribution may be considered a distribution for uncertainty in the parameter estimates.

Three simulation methods are employed to demonstrate the effect of assuming a bivariate normal distribution for parameters vs. perfect correlation, or independent parameters. Specifically:

(1) The slope and intercept of the regression line are described by a specific form of the bivariate normal distribution (i.e., follow *Steps 1, 2* in Exhibit B-9, and use Equation B-10 instead of *Step 4*).

(2) The slope and intercept of the regression line are described by a general form of the bivariate normal distribution (i.e., follow *Steps 1 to 4* in Exhibit B-9).

(3) The slope and intercept of the regression line are described by independent normal distributions (i.e., follow *Steps 1–4* in Exhibit B-9, but omit the correlation coefficient ρ in *Steps 2 and 4*).

For each approach, Monte Carlo simulations with I=5,000 iterations were run to determine the set of parameter values (β_0, β_1) for a simple linear regression equation. Typically, the uncertainty in the parameter estimates is not accounted for when simple linear regression equations are used to relate to exposure variables in a model. Such an approach may fail to account for important sources of parameter uncertainty. Figure B-10 (middle panel) illustrates the preferred approach for characterizing parameter uncertainty based on the bivariate normal distribution. (Note that the correlation coefficient relating the intercepts and slopes generated from the simulation is consistent with the correlation coefficient that describes the bivariate normal distribution; this is a good check that the simulation was set up correctly and run for a sufficient number of iterations). These results are contrasted with results using a form of the bivariate normal (Equation B-10) that underestimates uncertainty (top panel) unless parameters are perfectly correlated. In addition, the simplistic approach of sampling from independent normal distributions (bottom panel), yields a “shot gun” scatter plot. Sampling from independent normal distributions results in unlikely extreme combinations of the slope and intercept more often than the correct bivariate normal approach; propagating this bias through a risk model may severely bias estimates of uncertainty in risk.

Table B-9. Zinc concentrations in paired (i.e., co-located) soil and dust samples (ppm) for n=21 locations.

Sample	Soil (X_i)	Dust (Y_i)	Sample	Soil (X_i)	Dust (Y_i)
1	120	216	12	560	200
2	190	149	13	560	256
3	270	83	14	720	496
4	285	508	15	800	239
5	310	215	16	880	203
6	340	219	17	910	757
7	350	203	18	1035	676
8	380	101	19	1445	426
9	440	178	20	1600	522
10	480	232	21	1800	276
11	560	199			

Bivariate Normal Distribution for Parameters of the Regression Equation		
B_0	mean	173.9
	variance	4162.2
B_1	mean	0.193
	variance	0.0063
s^2		27857.4
Cov (B_0, B_1)		-4.2428
r		-0.8254

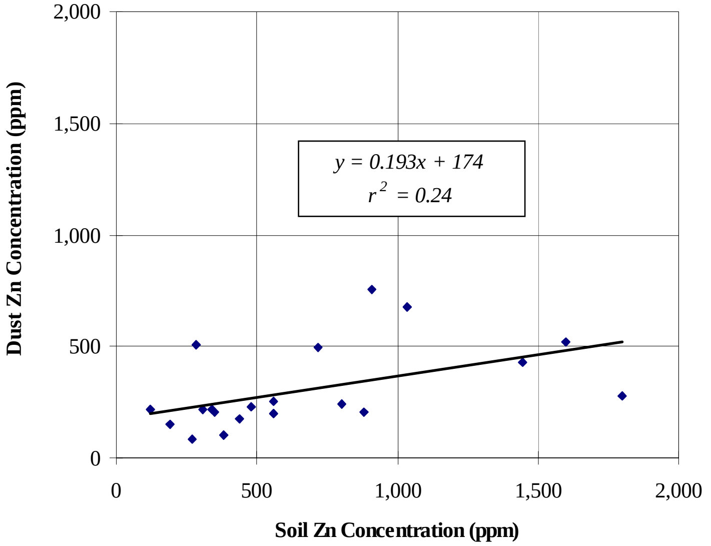


Figure B-9. Simple linear regression of zinc concentrations in soil and dust.

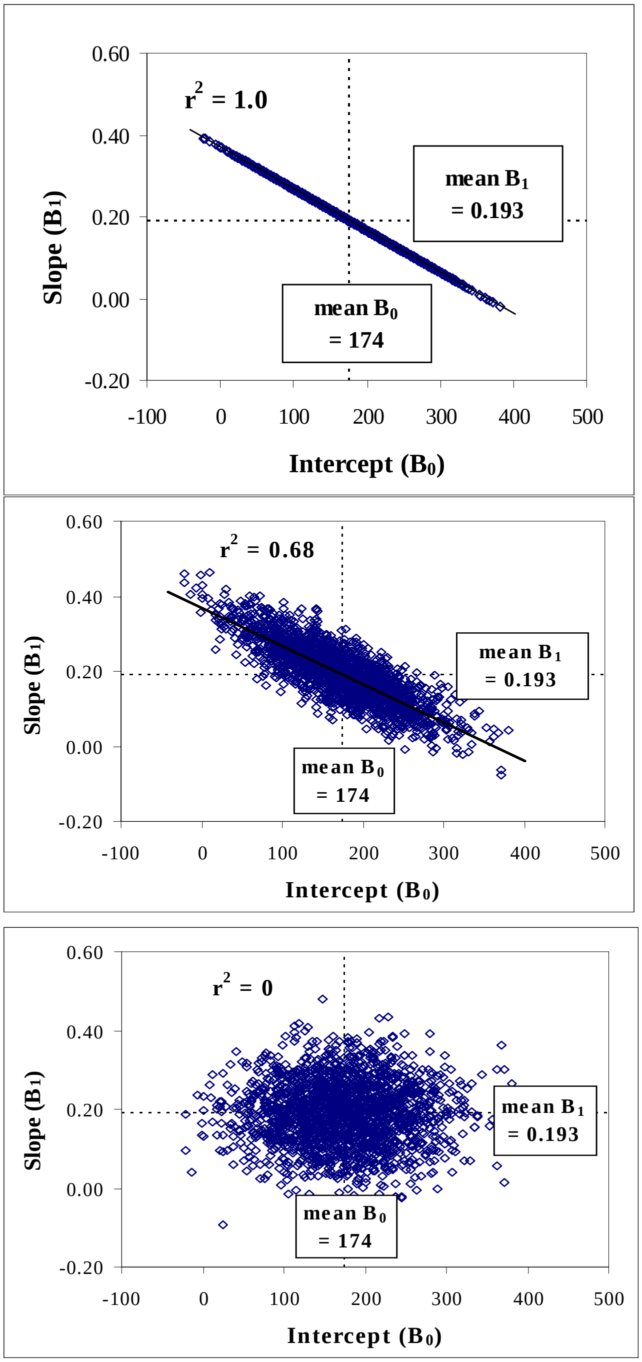


Figure B-10. Results of Monte Carlo simulation (*n*=5000 iterations) to estimate the slope and intercept of a regression equation. Top panel reflects the bivariate normal distribution for the special case that fails to capture the parameter uncertainty; middle panel reflects the preferred bivariate normal distribution with ρ=-0.825 based on empirical paired data; bottom panel reflects sampling from independent normal distributions.

REFERENCES FOR APPENDIX B

Brainard, J. and D.E. Burmaster. 1992. Bivariate Distributions for Height and Weight of Men and Women in the United States. *Risk Anal* 12(2):267–275.

Brately, P., B.L. Fox, and L.E. Schrage. 1987. *A Guide to Simulation*. Springer-Verlag, NY.

Burger, J., W. L. Stephens, Jr., C. S. Boring, M. Kuklinski, J.W. Gibbons, and M. Gochfeld. 1999. Factors in Exposure Assessment: Ethnic and Socioeconomic Differences in Fishing and Consumption of Fish Caught along the Savannah River. *Risk Anal*. 19(3):427–438.

Calabrese, E.J., Stanek, E.J., and Barnes R. 1996. Methodology to Estimate the Amount and Particle Size of Soil Ingested by Children: Implications for Exposure Assessment at Waste Sites. *Regul. Toxicol. Pharmacol*. 24:264–268.

Charney, E., J. Sayre, and M. Coulter. 1980. Increased Lead Absorption in Inner City Children: Where Does the Lead Come From? *Pediatrics* 65:226–231.

Conover, W.J. 1980. *Practical Nonparametric Statistics*. John Wiley & Sons, NY.

Cullen, A.C. and H.C. Frey. 1999. Probabilistic Techniques in Exposure Assessment. A Handbook for Dealing with Variability and Uncertainty in Models and Inputs. Plenum Press.

d'Agostino, R.B. and M.A. Stephens. 1986. *Goodness-of-fit techniques*. Marcel Dekker, Inc, NY.

Filliben, J.J. 1975. The Probability Plot Correlation Coefficient Test for Normality. *Technometrics* 17(1):111–117.

Finley, B.L. and D.J. Paustenbach. 1994. The Benefits of Probabilistic Exposure Assessment: Three Case Studies Involving Contaminated Air, Water and Soil. *Risk Anal* 14(1):53–73.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. Van Hostrand Reinhold, NY.

Gilliom, R.J. D.R. Helsel. 1986. Estimation of Distributional Parameters for Censored Trace Level Water Quality Data, 1. Estimation Techniques. *Water Resources Research*. 22:135–146..

Hahn, G.J. and S.S. Shapiro. 1967. *Statistical Models in Engineering*. John Wiley & Sons, NY.

Helsel, D.R. and R.M. Hirsch. 1992. *Statistical Methods in Water Resources*. Elsevier. Amsterdam.

Hoffman, F.O. and J.S. Hammonds. 1992. *An Introductory Guide to Uncertainty Analysis in Environmental and Health Risk Assessment.* ES/ER/TM–35. Martin Marietta.

Hora, S.C. 1992. Acquisition of Expert Judgment: Examples From Risk Assessment. *J. Energy Eng*. 118(2):136–148.

Johnson, N.L., S. Kotz, and N. Balakrishnan. 1995. *Continuous Univariate Distributions*. Volume 2, Second Ed. John Wiley & Sons, NY.

Law, A.M. and W.D. Kelton. 1991. *Simulation Modeling and Analysis*. McGraw-Hill, NY.

Looney, S.W. and T.R. Gulledge. 1985. Use of the Correlation Coefficient with Normal Probability Plots. *American Statist.* 39:297–303.

Mendenhall, W. and R.L. Scheaffer. 1973. *Mathematical Statistics with Applications*. Duxbury Press.

Mood, A.M. and F.A. Graybill. 1963. *Introduction to the Theory of Statistics*. Second Edition. McGraw-Hill, Inc.

Morgan, G.M. and M. Henrion. 1990. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, NY.

Nelsen, R.B. 1986. Properties of a One-Parameter Family of Bivariate Distributions with Specified Marginals. *Comm. Stat. (Theory and Methods)* 15:3277–3285.

Nelsen, R.B. 1987. Discrete Bivariate Distributions with Given Marginals and Correlation. *Comm. Stat. (Simulation and Computation)* B16:199–208.

Oregon DEQ. 1998. *Guidance for the Use of Probabilistic Analysis in Human Health Exposure Assessments*. Waste Management and Cleanup Division. Interim Final. November.

Ott, W.R. 1990. A Physical Explanation of the Lognormality of Pollutant Concentrations. *J. Air Waste Manage Assoc.* 40(10):1378–1383.

Ott, W.R. 1995. *Environmental Statistics and Data Analysis*. CRC Press, Boca Raton.

Palisade Corporation. 1994. *Risk Analysis and Simulation Add-In for Microsoft Excel or Lotus 1-2-3*. Windows Version Release 3.0 User's Guide, Palisade Corporation, Newfield, NY.

Roseberry, A.M. and D.E. Burmaster. 1992. Lognormal Distributions for Water Intake by Children and Adults. *Risk Anal.* 12(1):99–104.

Royston, J.P. 1982. An Extension of Shapiro and Wilk's *W* test for Normality to Large Samples. *Appl. Stat*. 31:115–124.

Snedecor, G.W. and W.G. Cochran. 1989. *Statistical Methods*. Eighth Edition. Iowa State University Press, Iowa.

Stanek, E.J. and Calabrese, E.J. 1995. Daily Estimates of Soil Ingestion in Children. *Environ. Health Perspect.* 103:176–285.

Thompson, K. 1999. Developing Univariate Distributions from Data for Risk Analysis. *Hum. Eco. Risk Assess.* 5(4):755–783.

Tukey, J.W. 1977. *Exploratory Data Analysis*. Addison-Wesley, Boston.

U.S. EPA. 1982. *Air Quality Criteria for Particulate Matter and Sulfur Oxides*. ECAO, ORD. EPA 600/8–82-029.

U.S. EPA. 1992. *Guidance for Data Useability in Risk Assessment, Part* A. Office of Emergency and Remedial Response, Washington, DC. OSWER Directive No. 9285.7-09A.

U.S. EPA. 1994. *Guidance for Conducting External Peer Review of Environmental Regulatory Models.* Office of the Administrator, Washington, DC. EPA/100/B-94-001. July.

U.S. EPA. 1997a. *Exposure Factors Handbook.* Office of Research and Development, Washington, DC. EPA/600/P-95/002Fa, Fb, and Fc.

U.S. EPA. 1997b. *Use of Probabilistic Techniques (Including Monte Carlo Analysis) in Risk Assessment,* Memorandum from Deputy Administrator Hansen and *Guiding Principles for Monte Carlo Analysis*. EPA/630/R-97-001.

U.S. EPA. 1999a. *Report of the Workshop on Selecting Input Distributions for Probabilistic Assessments.* Risk Assessment Forum. EPA/630/R-98/004. January.

U.S. EPA. 1999b. *Options for Development of Parametric Probability Distributions for Exposure Factors.* Office of Research and Development. Research Triangle Institute Final Report. April 6.

U.S. EPA. 2001. *Development and Evaluation of Probability Density Functions for a Set of Human Exposure Factors.* Office of Emergency and Remedial Response. University of California Draft Report. May.

Vose, D. 1996. *Quantitative Risk Analysis: A Guide to Monte Carlo Modeling*. John Wiley & Sons, NY.

Wonnacott and Wonnacott. 1981. *Regression: A Second Course in Statistics*. John Wiley & Sons, NY.